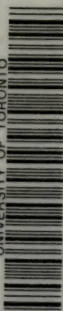


UNIVERSITY OF TORONTO



3 1761 00458024 7

UNIV. OF
TORONTO
LIBRARY

~~RETURN TO
DEPARTMENT OF PSYCHOLOGY LIBRARY
UNIVERSITY OF TORONTO~~



Digitized by the Internet Archive
in 2007 with funding from
Microsoft Corporation

(PSYCHOLOGICAL REVIEW PUBLICATIONS)

Psychological Monographs

EDITED BY

JAMES ROWLAND ANGELL, UNIVERSITY OF CHICAGO
HOWARD C. WARREN, PRINCETON UNIVERSITY (*Review*)
JOHN B. WATSON, JOHNS HOPKINS UNIVERSITY (*J. of Exp. Psych.*)
SHEPHERD I. FRANZ, GOVT. HOSP. FOR INSANE (*Bulletin*) and
MADISON BENTLEY, UNIVERSITY OF ILLINOIS (*Index*)

VOLUME XXIV

1917



146167

5 | 6 | 18

PSYCHOLOGICAL REVIEW COMPANY

PRINCETON, N. J.

AND LANCASTER, PA.

AGENTS: G. E. STECHERT & CO., LONDON (2 Star Yard, Carey St., W. C.)
LEIPZIG (Koenigstr., 37); PARIS (16 rue de Condé)



BF
1
P69
v.24

111

TABLE OF CONTENTS OF VOL. XXIV

1. Two Studies in Mental Tests. I. Variable Factors in the Binet Tests. II. The Diagnostic Value of Some Mental Tests. CARL C. BRIGHAM. Pp. 254.
2. Radiometric Apparatus for Use in Psychological and Physiological Optics. C. E. FERREE and GERTRUDE RAND. Pp. xvi + 65.
3. Studies from the Psychological Laboratory of the University of Chicago. Transfer of Training and Retroaction. LOUIE WINFIELD WEBB. Pp. 90.
4. Studies from the Psychological Laboratory of the University of Chicago. The Reliability of Mental Tests in the Division of an Academic Group. BEARDSLEY RUMML. Pp. 63.
5. The Analysis of Mental Functions. CURT ROSENOW. Pp. 43.

(This volume includes Monographs 102-106)

Vol. XXIV
No. 1

PSYCHOLOGICAL REVIEW PUBLICATIONS

Whole No. 102
1917

THE Psychological Monographs

EDITED BY

JAMES ROWLAND ANGELL, UNIVERSITY OF CHICAGO

HOWARD C. WARREN, PRINCETON UNIVERSITY (*Review*)

JOHN B. WATSON, JOHNS HOPKINS UNIVERSITY (*J. of Exp. Psychol.*)

SHEPHERD I. FRANZ, GOVT. HOSP. FOR INSANE (*Bulletin*) and

MADISON BENTLEY, UNIVERSITY OF ILLINOIS (*Index*)

Two Studies in Mental Tests

I. Variable Factors in the Binet Tests

II. The Diagnostic Value of Some Mental Tests

BY

CARL C. BRIGHAM, PH.D.

Instructor in Psychology
Princeton University

PRINCETON CONTRIBUTIONS TO PSYCHOLOGY

PSYCHOLOGICAL REVIEW COMPANY

PRINCETON, N. J.
AND LANCASTER, PA.

AGENTS: G. E. STECHERT & CO., LONDON (2 Star Yard, Carey St., W. C.)
LEIPZIG (Hospital St., 10): PARIS (76 rue de Rennes)



ACKNOWLEDGMENTS

The writer is indebted to Prof. H. C. McComas, Mr. Prentice Reeves and Mr. Norbert J. Melville for their cooperation in examining the Princeton children. The four experimenters are indebted to Miss Mabel T. Vanderbilt, supervising principal of the Princeton schools, not merely for allowing the use of the subjects, but for her careful attention to the administrative details of the experiment which facilitated the testing and made the work a source of pleasure.

The writer is indebted to Superintendent E. Mackey for the privilege of working in the Trenton schools, and wishes to express his deep gratitude to Dr. J. M. McCallie, supervisor of special classes and principal of the Franklin school, who willingly sacrificed much of his time in making all the arrangements for the conduct of the Trenton experiment. Without the cooperation of Dr. McCallie, the principals of the Carroll Robbins, Columbus, Cooper, Girard, U. S. Grant, and Jefferson schools, and the teachers of the special classes, the Trenton experiment would have been impossible.

Both the Princeton and Trenton investigations were carried out under the direction of Prof. H. C. Warren and Prof. H. C. McComas who at all times gave their expert guidance in devising and conducting the experiments, in suggesting methods of treating the data, and in correcting and revising the manuscript. The writer is deeply indebted to them for their assistance.

VARIABLE FACTORS IN THE BINET TESTS

TABLE OF CONTENTS

I.	Introduction	I
II.	Subjects and Methods	8
III.	The Personal Equation	18
VI.	Grade Correlations	37
V.	Sex Differences	65
VI.	Summary	91

I. INTRODUCTION

During the past decade, the Binet-Simon measuring scale for intelligence has received considerable attention, and a large amount of literature has appeared on the subject. No attempt has been made in the following pages to review all the literature on this scale or other systems of intelligence testing. Kite (38) gives an excellent account of the history and nature of the scale. Kohs (41) has assembled a very complete bibliography on the subject up to June 1914. Schmitt (57) gives an historical account of the development of the various attempts to correlate psychological findings with general intelligence, particularly in this country and England. Bobertag (10) and Schmitt both give detailed descriptions and analyses of the individual tests. Stern (62) has devoted a monograph to the collection, exposition and critical analysis of the large amount of data bearing on the problem of intelligence testing, and in another work (61) has assembled the literature of cognate fields. The literature bearing on the Binet scale up to 1912 is largely descriptive of the scale itself, the standard methods of procedure, etc. The more recent literature has been critical and reveals a tendency at the present time for investigators to depart from the methods of the extensive application of the scale as a whole to the more intensive study of the individual tests.

All systems of intelligence tests may be classified as qualitative or quantitative. The qualitative system consists of an aggregation of tests designed to detect the capacities or incapacities of the subject in order to afford the experimenter an opportunity to make a diagnosis concerning the subject's mentality. This method throws the responsibility for the final diagnosis on the experimenter. The system of tests proposed by Healy and Fernald (34) are of this type. Quantitative systems of tests necessitate a final score of some sort, whether that score be in the form of a mental age, a mental quotient, a certain number of points,

a coefficient of intellectual ability, a percentile rank or what not. The essential characteristics of the quantitative systems are the interpretation of the total scores in terms of the age of the subject, and the placing of the responsibility for the final diagnosis on the tests rather than the experimenter.

Binet and Simon's 1905 scale (5 and 6) was of the qualitative type. A series of 30 tests of approximately increasing difficulty was published with directions for their application. The authors reported in a general way that from their experience in examining a few selected normal children of different ages, and other subnormal children in the schools and at the Salpêtrière, approximate levels of performance could be found characteristic of the development of normal children of 3, 7, 9 and 11 years chronologically, the performance of idiots, imbeciles and morons corresponding roughly with that of normal children of 3, 7 and 9. Although the reference to chronological ages introduced the quantitative element, at no place were the authors insistent on this point, merely stating that they had found the series of tests exceedingly valuable in diagnosing and classifying defectives, and in their opinion others would also find it valuable.

The 1908 scale (7) was quantitative in character owing to the introduction of the concept of "mental age". It included a list of 56 tests grouped according to ages from 3 to 13, each group containing from four to eight tests. Most of the tests of the 1905 series were included, the additions including in a large measure tests of a scholastic nature. The authors gave directions for applying the series and for computing the resultant "mental age". A child testing three years below his chronological age was to be considered defective.

Although the scheme of the 1908 series was entirely quantitative, the authors did not discard the qualitative idea, and they cautioned against the application of the scale in the manner of a measure of height or weight. The border line between the idiot and the imbecile was fixed by the ability to use and comprehend spoken language. The imbecile was differentiated from the moron by the use of written language, illiteracy being di-

ferentiated from imbecility by certain tests. The authors stated that the moron could be defined only in terms of the environment in which he lived, and they considered six tests important in differentiating the moron from the normal individual of the Paris population. Any system of tests which throws more weight on some tests than on others in making a differential diagnosis is fundamentally qualitative in kind, for the responsibility is placed not on the score but on the judgment of the experimenter. The idea of a quantitative measuring scale of intelligence however met with instant favor. The interest that actuated the psychologists of the "early nineties" to correlate the measurements of reaction time, motor ability, sensory discrimination, etc. with intelligence was revived. The scale was translated into several languages and applied to individuals of many classes and types.

In 1911, the authors published a revised scale (8) in which many of the tests of scholastic ability were discarded, and the remaining tests shifted about so that there were five tests for every year except one from III to X with similar groups for "twelve year", "fifteen year" and "adult" mentality. In the same year, Binet published an article (4), his last word on the subject, in which he discussed many of the criticisms which the scale had received, and again sounded the note of warning against the mechanical interpretation of results. However, as one traces Binet's thought on the subject through his writings, he may see the idea of a qualitative system of tests gradually dropping into the background, and more and more weight placed on the "scientific" (quantitative) measure of intelligence.

That Binet did not depart entirely from the qualitative standpoint is shown by his discussion of the test of comprehending difficult questions. "Sometimes after an examination one hesitates on a diagnosis. The child has failed in one or two tests, but this does not seem to be convincing. Failure to give the day and date and the months of the year are excusable errors, which may be caused by distraction or by lack of education. But the questions for comprehension dissipate all doubts. We recall

several instances when teachers brought us children, desiring to know whether or not they were abnormal; occasionally, in this way they set a trap for us, but we did not object, it was fair play. Our questions for comprehension decided us every time. We remember one child who was very slow in answering as though dull, his face was expressionless and unprepossessing; he knew neither the day nor the date, nor what day comes after Sunday, and he was $10\frac{1}{2}$ years old; his reading was syllabic. But when we asked question 5: Why do we judge a person by his acts rather than by his words? he gave the following answer: Because words are not very sure and acts are more sure. This was enough—our opinion was formed, that child was not so bad as he seemed." (Town's (72) translation, page 48.)

The popular interest that was manifest before the advent of the 1911 scale was tremendously reinforced in this country by Goddard's (30) publication of the results of the application of the scale to "two thousand" non-selected school children in Vineland, N. J. Popular interest increased rapidly, and the scale continued to have wider and wider application in the hands of less and less experienced investigators. The concept of "mental age" was exceedingly easy of comprehension, no apparatus was needed, and the scale has now become the common property of all. This development or overdevelopment has taken place in spite of the warnings of the authors themselves and the psychological fraternity in general. The very fact of overdevelopment however is striking evidence that persons interested in the social sciences need a quantitative scale for measuring intelligence.

The question whether the Binet scale is an accurate measure of intelligence can be decided only by the study of the individual tests and the factors underlying them. A study of this sort will show the errors that underlie the total score or "mental age", and at the same time will show the direction in which the correction of the scale should take place. The proper understanding of the individual tests involves the theory on which the measuring scale was constructed.

The method which Binet and Simon used in constructing their

measuring scale of intelligence was entirely empirical. A large number of tests were given to children of a certain social status. Certain tests could be shown to be correlated with age, and in the authors' opinion were correlated with intelligence. The fact that at a certain age a test could be passed by a certain proportion of the subjects was taken to mean that the test in question was characteristic of that age. Tests that were characteristic of the same age level were then combined into one age group. In this way a scale was built up with a number of tests for each age group. By a certain arbitrary system of scoring the reactions of a subject to all or part of the scale of tests, the "mental age" of the subject was obtained. The comparison of the "mental age" with the chronological age of the subject would show him to be advanced, at age or retarded, and the amount of acceleration or retardation would afford a quantitative index of his intelligence.

A person could construct a scale on the same basis and arrive at an age score using entirely different tests. A scale could be constructed containing tests of height, weight, vital capacity, strength of grip, circumference of the head, etc. and the results interpreted in terms of age. In this case however the age obtained would be more physical than mental. A scale of tests could also be constructed which involved the subject's knowledge of geography, spelling, history, grammar, etc. but in this case the resulting age would be determined very largely by the amount of training the subject had received.

The assumptions that a child at a certain age should weigh 25 pounds, at another age 50 pounds, etc., that a child can repeat 3 digits at one age, 5 digits at another and 7 digits at another, and that a certain percentage of children at one age can enumerate the months, and a higher percentage at another age, differ only in the possible determiners to which the growth may be referred. In the first case the growth is referred to certain physiological processes which are supposedly independent of intelligence and training. Binet believed that the principal determiner of growth in the last two cases was intelligence, but the possibility

remains that they might be more or less independent of intelligence, and more or less dependent on training and other variable factors.

The principle on which the scale was constructed involves three assumptions, (1) that the individual tests are correlated with age, (2) that the individual tests are correlated with intelligence, and (3) that intelligence is correlated with age—three distinct assumptions any one of which does not necessarily involve the others. The purpose of this investigation is to study the correlation of the individual tests with age, to determine the variable factors that might operate on the tests to produce an apparent correlation with age that was not a real correlation, or that might alter the real correlation in some way.

There is a possibility that an error might occur in the statistical treatment of the results, so that figures which would apparently indicate a correlation with age of a certain degree might actually represent a correlation of another degree. Another variable factor is the personal equation of the experimenter, who might alter the procedure in giving a certain test so that the correlation of that test with age might be different from the correlation obtained by another experimenter. If the subjects of various ages had received different school training, this difference might introduce another factor which would vary independently of the age of the subjects. If the tests used depended on any inherited or acquired differences between the sexes, then the correlation of the tests with age might be different for the two sexes. If any or all of the variable factors mentioned prove to be present in the correlation of the tests with age, then certain allowances will have to be made for these factors in making a diagnosis of the subject's intellectual ability on the basis of his total score or "mental age", and the scale becomes qualitative rather than quantitative.

At the Fourth International Conference for School Hygiene held in Buffalo in the summer of 1913, several persons of unquestioned authority in the field of mental tests held an informal

conference on the Binet-Simon scale, reporting the results in 1914 in the form of recommendations and suggestions (15). The question, "How much is the outcome of the testing influenced by the personal equation, both of the examiner and examinee?" was answered, "Undoubtedly there is some influence and it may be a serious source of error." Another question, "How much do previous environment and school training effect the outcome of the tests?" was left unanswered by the opinion, "The experimental evidence thus far available is conflicting. Further investigation is needed." The question, "Should the scale be divided, in the upper years at least, to furnish separate standards or separate tests for the two sexes?" was answered, "We do not know, and recommend this a subject for investigation." The following study is in part an attempt to answer these questions.

The method used in this study is that of studying the individual tests, disregarding entirely the total score or "mental age". There are at present so many revisions and editions of the Binet scale, that the term "mental age" has no meaning outside of the particular scale in question. The tests that are used in the various standardizations are however approximately the same, so that conclusions concerning the factors underlying the individual tests have a wider significance than those drawn from the "mental ages". Furthermore variable factors in the individual tests may balance each other in the total score so that their influence might be obscured.

The subjects and methods will be described first, and in connection with the methods of treating the results a statistical error will be pointed out. The problems of the personal equation, grade correlations and sex differences will then be taken up in detail.

II. SUBJECTS AND METHODS

SUBJECTS

The data which are here analysed to determine the influence of the personal equation, of grade training and of sex differences, are derived from all the boys and girls below the seventh grade in the Princeton, N. J., Model School. This group includes 422 subjects of the following age distribution,—

CHRONOLOGICAL AGES.

4	5	6	7	8	9	10	11	12	13	14	15	16
4	17	62	52	56	42	53	49	36	32	11	6	2

Each of the first six school grades was divided into a plus and minus grade, the latter division being under a different teacher, and containing those who were either backward, or, on account of illness, change of school, or for reasons not necessarily related to their mental development, were not sufficiently advanced to perform the work of their grade. The school also contained a special class for defective and exceptionally backward children. The subjects were distributed in the school grades as follows,—

SCHOOL GRADES.

Spec.	Kind.	I—	I+	II—	II+	III—	III+	IV—	IV+	V—	V+	VI—	VI+
18	32	38	51	12	40	12	45	15	35	15	49	11	49

39 or 9.2% of the subjects were children of non-English speaking parents, this group including 6.6% of the children in the Kindergarten and first six regular grades, and 15.7% of those in the special class and minus grades.

The selection of subjects is only fairly typical of the general run, for Princeton has no manufactories. The children examined came, for the most part, from the homes of laborers, domestics, artisans, farmers, tradesmen, clergymen and college professors. The selection is atypical in that none of the children came from homes of the manufacturing class, while an unusually large pro-

portion came from the homes of those engaged in domestic, personal, and professional service.

TESTS

The scale used was Goddard's (28) 1911 revision of the Binet-Simon scale. The methods used in giving the tests were, as far as possible, the same as those outlined by Goddard in the original revision, incorporating the rules and suggestions for standardized scoring published by that writer (29) in 1913. The methods used will not be discussed in detail, for the data are not used in obtaining age norms and standards for children generally. For the analysis of the data in terms of grade and sex it is not necessary that the procedure should be absolutely standardized, but that the experimenters who gave the tests should have used the same procedure. Differences in the technique of the experimenters will be discussed in the chapter on the personal equation.

One variation from the usual procedure was adopted. In no case did the experimenter know the chronological age of the child being tested. The influence of any prejudice or bias on the part of the experimenter is therefore eliminated from the problem of the correlation of the tests with age. The three experimenters who gathered the material in the spring of 1913 examined the sixth grade first and the remaining grades in decreasing order. During the school year 1913-1914, the fourth experimenter examined all children at that time in the kindergarten and first grades, and others who were not examined in the spring of 1913.

The tests in the "three year", "four year", "five year", "fifteen year" and "adult" groups were given so infrequently that the data from them are not treated. The tests used are as follows. The figure at the right shows the total number of times each test was given.

AGE VI

1. Distinguishing between morning and afternoon.....	108
2. Defining in terms of use.....	333
3. Executing three commissions.....	100

4. Showing right hand and left ear.....	107
5. Choosing the prettier of given faces.....	117

AGE VII

1. Counting 13 pennies.....	217
2. Describing pictures.....	219
3. Indicating omissions in pictures.....	217
4. Copying the diamond (in pencil).....	225
5. Naming four colors.....	218

AGE VIII

1. Comparing remembered objects (butterfly and fly).....	271
2. Counting backwards from 20 to 0.....	251
3. Enumerating the days of the week.....	277
4. Counting stamps.....	258
5. Repeating 5 digits.....	413

AGE IX

1. Making change.....	271
2. Defining in terms superior to use.....	333
3. Giving the day and date.....	307
4. Enumerating the months.....	284
5. Arranging five weights.....	334

AGE X

1. Recognizing pieces of money.....	282
2. Copying designs from memory.....	252
3. Repeating 6 digits.....	413
4. Comprehending easy and difficult questions.....	250
5. Using three words in sentence (two ideas).....	279

AGE XI

1. Detecting absurdities in statements.....	226
2. Using three words in sentence (one idea).....	279
3. Giving 60 words in three minutes.....	233
4. Giving rhymes with day, mill and spring.....	213
5. Reconstructing dissected sentences.....	190

AGE XII

1. Repeating 7 digits.....	413
2. Defining abstract terms.....	144
3. Repeating a sentence of 28 syllables.....	169
4. Resisting suggestion (length of lines).....	203
5. Solving problems from various facts.....	123

The tests in the "six year" group, with the exception of defining in terms of use, and the tests in the "twelve year" group, with the exception of repeating 7 digits, were given so infrequently or so irregularly that the data from them could not be treated. The apparatus used in the test of arranging five weights was not constant throughout the experiment, the standard cubes

and weighted pill boxes being used at different times by different experimenters. On this account, the data from this test are not included in the subsequent discussion.

METHODS OF TREATING RESULTS

The chronological age of each subject was taken as that at the last birthday, one tenth of a year being allowed for each 36 days beyond the birthday. The subject that was 10 years and 35 days would be rated 10.0 years, while ten years and 36 days would be 10.1 years. A subject one day short of 11 would be rated 10.9 etc. The teachers of each grade submitted the dates of birth of all pupils after the grade had been tested. These data were later checked up from the entrance cards. Since the purpose of this study is to analyze the factors involved in the individual tests, no "mental ages" or total scores were figured. The classifications of the subjects are all made independently of the tests.

Two measures of central tendency will be used in the subsequent discussion, the average and the median. The measure of variability from the average, that will be used, is the mean variation (or average deviation), the average of the differences, regardless of signs, between the separate measures in the series and the average of the whole series. The measure of variability from the median that will be used is the semi-interquartile range (Q), or half the difference between the measure with three times as many measures above as below it and the measure with one third as many measures above as below it, i. e. half the difference between the 25 percentile, and the 75 percentile. Any coefficients of correlation used will be stated in terms of the formula applied. The reader is referred to Thorndike (70) for the discussion and explanation of the statistical measures used.

The measures of ability in most of the tests are in the "all or none" form—the tests are either passed or failed. The only measure that can be obtained from data of this sort is the percentage that an ability is present in a defined group. This method of treating the results has as many "pit-falls" as the tests themselves. Before undertaking the analysis of the Prince-

ton data to determine the effect of the personal equation of the experimenter, and the age, grade, and sex of the subject upon the results of the individual tests, it is necessary to consider an error which underlies incomplete data, or those data derived from experimenting in which every test is not given to every subject.

No uniform instructions were given to the experimenters concerning the order in which the tests should be given, nor the number of tests that should be tried. The experimenters attempted to determine the mental age of the child according to the scale. In doing this they would start with some test which they considered would be interesting to the child, and, at the same time, well within his reach. The tests given first were usually those of describing pictures and arranging five weights. The experimenter would then gradually explore the subject's range of ability, varying the order of the tests so as to maintain the subject's interest, and to ward off fatigue. In this way the experimenter would eventually establish the basal age of the subject (that age in which he passed all five of the tests), and by the end of the examination would have tried all the tests above the basal age which, in his judgment, there was any possibility of the subject's passing. This method of experimenting will be called incomplete. The other method of experimenting, in which a certain number of tests are adopted and all of the tests are tried on each subject, will be called complete. Each experimenter in the Princeton investigation averaged 19 or 20 tests to a subject. In the Trenton investigation all the tests were given to all the subjects.

The incomplete method is more desirable from the standpoint of the subject who is not unnecessarily fatigued, and from the standpoint of the experimenter, as well, who saves in the expenditure of time and energy. However, the data derived from the incomplete method are subject to an error, which, unless it is properly considered, will completely vitiate the results.

When the experimenter does not try a test above the basal age because he believes that the subject will not pass it, he implies that the subject will fail it. This amounts to a failure,

for the subject receives no credit. However, a failure of this sort, due to the experimenter's assumption, is not the same as an actual failure in which the test is tried, for there is always the possibility that the assumption was unjustified. In like manner when the experimenter does not try tests below the basal age, he actually gives credit for passing the test without the actual trial.

In some cases the assumption on the part of the experimenter is quite justified. Obviously if a subject can make change, he can count up to thirteen; if he can repeat seven digits, he can repeat five and six digits; if he knows the names of the months, he will know the days of the week; and, conversely, if he cannot repeat the days of the week, he cannot repeat the months. Other assumptions are less justifiable. Since very intelligent persons, lacking in particular sorts of abilities, might fail in tests such as drawing the design from memory or arranging five weights, there is no reason for supposing that a subject making basal "eleven" or "twelve" will pass these tests. At the same time there is no reason for assuming that a subject failing to establish basal "seven" for instance, will fail to pass a test such as the line suggestion test in "twelve". The assumptions of the experimenters, then, are more or less justifiable and it is impossible to estimate the amount of the justification, since this is dependent on the nature of the individual tests.

The manner in which this error works out in the statistical treatment of the results may be shown by examining any test which has been tried through a number of chronological ages. Table I shows the results of the 60 word test obtained from subjects 7 to 13 years of age.

TABLE NO. 1

Analysis of the Results from the Test in Naming 60 Words in 3 Minutes.

Chronological ages	7	8	9	10	11	12	13
No. of times given.....	11	18	25	42	44	31	28
No. of time passed	4	10	10	24	34	19	21
Actual percentage passed.....	36%	56%	40%	57%	77%	61%	75%
Total number of subjects.....	60	52	42	54	48	36	28
Percentage of subjects to whom test was given.....	18%	35%	60%	78%	92%	86%	100%
Theoretical percentage passed....	7%	19%	24%	44%	71%	53%	75%

An example will make the above table clear. The 60 word test was given to 11 subjects, age seven, 4 of whom passed. In all there were 60 subjects at this age, so that the 11 subjects to whom the test was given constitute but 18% (and probably the brightest 18%) of this whole number. The percentage passed would have been 7% had the test been given to all 60 subjects, and had all the subjects failed who the experimenters assumed would fail if they gave the test. The true per cent. which represents the ability of non-selected seven year boys and girls in passing the 60 word test therefore lies somewhere between 7% and 36%, probably nearer 7%. An accurate estimate of the real per cent. which will represent this ability is, however, impossible. In like manner, the ability of the 8 year subjects is represented by a percentage somewhere between 19% and 56%.

As the proportion between the number of subjects in the group and the number actually tested increases, the disparity between the actual and theoretical percentage passed becomes less, or, in other words, the results which express the ability of a group become more reliable as the number of individuals actually tested as a sample of this group becomes larger. The higher the percentage given, the more reliable the percentage passed, when the reliability is measured by the difference between the actual percentage passed and the theoretical percentage passed.

The source of error mentioned causes great difficulty in comparing the results of different investigators. For example, it is desired to compare the results of Terman and Childs (66) and Dougherty (23) with those of this investigation on the 60 word test. Table 2, derived from their published results, shows the percentage that the test was given of the number of times it was possible to be given, (%G), the actual percentage passed, (A%P), and the theoretical percentage passed, (T%P), or that percentage passed that would have resulted had all of the subjects failed, who it is necessary to suppose would have failed, had the test been given all the possible number of times.

TABLE NO. 2

Analysis of the Results of Three Investigators on the 60 Word Test.

Age	This investigation			Terman and Childs			Dougherty		
	%G	A%P	T%P	%G	A%P	T%P	%G	A%P	T%P
7	18	36	7	14	50	7			
8	35	56	19	47	35	16	15	0	0
9	60	40	24	86	57	49	35	60	21
10	78	57	44	100	67		78	53	41
11	92	77	71	98	83	82	89	79	70
12	86	61	53	97	82	80	91	95	87
13	100	75		100	94		94	88	83

It is very difficult, if not impossible, to make a comparison of these results shown in Table 2 for the years 7, 8 and 9. The ability of Terman's 7 year group is represented by a figure somewhere between 7% and 50%, while that of the 8 year group falls somewhere between 16% and 35%. Dougherty's 9 year group falls between 21% and 60%. In the older years where the results have greater reliability, it is probable that the discrepancies between the investigators could be accounted for on the basis of the inferiority of the selection of the older subjects in this investigation, the other investigations including children from the seventh and eighth grades.

In order to make a comparison between investigators, it is necessary to express the results in terms of a percentage or a proportion. The expression of the ability of a group by a percentage or a proportion is inaccurate if the data are incomplete, and in order to judge the accuracy of the data, it is necessary to know the degree of completeness. Unfortunately, the results of most of the investigations on the individual tests are not published in a form that enables one to estimate the accuracy of the data. The writers who have published their data in a form that will admit of this treatment, have not treated the sexes separately. On this account, the writer will not attempt a systematic comparison of the results of this investigation with those of other experimenters.

Before analysing the Princeton data the following problem should be answered:—What proportion of a given group must actually be tested for an ability in order that the results may be

considered as typical of the ability of the whole group? The proper proportion to select as typical of any one group depends upon the characteristics of the group itself. If the members of a group are similar, a smaller proportion would stand for the ability of the group than would be necessary for a group composed of unlike individuals. A smaller number of individuals would be necessary to stand for the ability of all the 12 year boys in the sixth grade, for example, than for all the 12 year boys coming from a great many grades. This proposition operates directly counter to actual practice, for the members of a group of similar individuals will be given similar tests, while unlike individuals will receive different tests, inasmuch as the experimenter adapts his procedure to the need of the individual being examined. The proposition actually means, then, that selected results from incomplete testing are more reliable than non-selected results, if each group has the same range of testing. The proportion of a group that must be tested to stand for the whole group will also vary from test to test. In some tests of particular abilities, no proportion will accurately stand for the whole group—the entire group must be tested. In other tests that are easy for the group, the results of a very small proportion would not be altered by examining the remainder of the group.

The problem of deciding what proportion of a given group must actually be tested for an ability in order that their results may be considered as typical of the ability of the whole group has, therefore, no answer in the work. The writer will decide arbitrarily what the proportion will be. The actual magnitude of the proportion between the number actually tested and the number in the whole group (the percentage given) will always be published as an index of the reliability of the percentage that the group passes the test in question.

It is not possible to obtain reliable results showing the growth of an ability with age, if the data on which the results are based are of the incomplete sort. A test for any age will be given to a superior selection of subjects below that age, and an inferior selection of subjects above that age, so that the growth curve

will appear flatter than it actually is. For this reason, the Princeton data may not be used for the purpose of standardizing age norms.

Binet (4) recognized the fallacy of calculating proportions from the actual number of times a test was given and passed when the test had not been given all the possible number of times. In calculating the proportions from Levistre and Morle's data, Binet used what the present writer would call the "theoretical proportion passed".

It has been shown that the reliability of the theoretical percentage passed rests on the accuracy of the experimenters' assumptions, and that according to the nature of the tests and the character of the groups to which they are given these assumptions vary from complete certainty to absolute uncertainty. Inasmuch as these assumptions are not equally certain, the conclusions drawn from them are not equally certain, and the logic of scientific method demands that an investigator establish the degree of certainty of his conclusions. In this case the measure of the degree of certainty is the magnitude of the percentage given.

The use of the theoretical percentage passed without reference to the percentage given ignores the dictum that an investigator establish the degree of certainty of his conclusions, and sets up all conclusions as equally valid, a procedure which in actual practice results in making all conclusions equally invalid when the fact of degrees of certainty is admitted. The investigator who draws conclusions from incomplete data should always state the percentage given and the actual percentage passed. This much at least is experiment. The only legitimate use of the theoretical percentage passed is when it is compared with the actual percentage passed as a probable limiting value. The theoretical percentage passed alone has no claim to reliability.

III. THE PERSONAL EQUATION

Before attempting to correlate the individual tests with age, grade and sex, it is necessary to demonstrate the presence or absence of the effect of the personal equation. By the term "personal equation" is meant the complex of variable factors which are independent of the mental make-up of the subject and the environmental conditions at the time of the examination. The term includes such widely different factors as the experimenter's ability to obtain the cooperation of the subject, his procedure in giving the tests, his criteria in deciding whether a subject's response should pass or fail, and the tests used, insofar as the selection of tests and the construction of the apparatus were occasionally left to his discretion, apart from the uniform procedure.

The only method of detecting the influence of the personal equation in most of the tests is that in which the responses of similar groups of subjects to different experimenters are compared. On account of the wide variations in the character of the subjects examined, it is not possible to compare similar groups. On some tests, however, it is possible to determine the effect of the personal equation independently of the method of group comparison. The results of the tests that may be studied independently will be discussed at some length, in order to demonstrate the fact that certain tests are susceptible to this influence.

The examinations of the Princeton subjects were made by four experimenters, called for convenience A, B, C and D. None of the experimenters was highly trained in giving the tests, although they had all been trained in the methods of psychological experimentation, one experimenter being an assistant professor of psychology, and the other three graduate students of psychology of at least one year's standing. B, C and D performed their experiments at the same time, in the spring of 1913, while A experimented one year later. B, C and D studied

the scale together so that it was possible to secure a correspondence in method. At the close of practically every day's testing, B, C and D would confer on the questions brought out by the day's work, and as far as possible would adopt uniform methods of procedure and scoring. A was subsequently trained in these same methods.

In spite of the attempt to adopt uniform methods, there were a few tests which always caused difficulty, and concerning which the experimenters could reach no definite agreement. One of the tests that caused the greatest difficulty was that of defining in terms of use and in terms superior to use. The hierarchy of responses to this test could be fairly arranged as follows. To the question "What is a chair?" the following typical responses would be obtained,—1, "A chair is a chair." 2, "This is a chair." 3, "A chair is to sit on." 4, "A chair is what you sit on." 5, "A chair is a thing you sit on." 6, "A chair is a piece of furniture you sit on." 7, "A chair has four legs, a back, etc." 8, "A chair is a piece of furniture with four legs, a back, etc." Any of the objects for which a definition is asked (fork, table, chair, horse, mother) may be defined by repetition, by demonstration, by indicating the use to which it is put, by showing the class to which it belongs, by describing its parts, or by the combination of any or all of these methods.

The only problem is to decide, arbitrarily, how definitely the class must be indicated (i. e. by "what", "a thing" or "a piece of furniture") in order to have the definition considered as one of classification. The rule adopted in this study was to consider "thing" as indicating the class. Nos. 1 and 2, definitions by repetition and demonstration, received no credit in "six years". Nos. 3 and 4 were given credit in "six years" as definitions by use, and nos. 5, 6, 7, and 8 were given credit in "nine years" as definitions in terms superior to use.

In studying the ranks given to the responses of the subjects in this test, it was found that the experimenters did not record the responses all of the time. A gave the test 94 times, and recorded the responses 66% of this number. B gave the test 98 times, recording the subject's answer 67% of the time. C

gave the test 65 times, and recorded the answer in 95% of the cases, while D gave the test 76 times and recorded the response only once.

By ranking the recorded responses of A, B and C according to the rules shown above, it is possible to obtain an estimate of the relative severity of their criteria in marking these responses plus or minus. 19% of A's definitions were corrected, the correction in all cases being from minus to plus. 11% of B's definitions were corrected, all of the corrections being from plus to minus. 17% of C's definitions were corrected, three fourths of them being changed from plus to minus, and one fourth from minus to plus. C's standards changed during the course of the experiment, so that at first, with older subjects, he was too lenient, while later, with younger subjects he was slightly too severe. The tendencies of A and B remain constant throughout the experiment, A marking too severely and B slightly too leniently. The differences between the experimenters hold constant for both sexes. The experimenters agreed on all definitions by use, the cases of disagreement coming on the definitions superior to use.

One test in which variations between the experimenters might be expected is that of copying the diamond. In this test, although the apparatus and procedure were the same, the experimenters had very little to guide them in forming their judgments of passed and failed. The instructions given ("The result is considered satisfactory if it would be recognized as intended for a diamond shaped figure"), and the examples published furnish very vague criteria.

In order to determine the effect of the personal equation of the experimenters in giving credit on this test, all of the reproductions of the diamond obtained in the Princeton and Trenton experimenting, (311 in number), were first transcribed and then ranked. On the sheet containing the copy only the subject's number was placed, so that the person ranking the reproductions was in ignorance of the experimenter by whom it was obtained, the mark that the experimenter had given it, and the age, grade, sex, etc. of the subject. The 311 diamonds were then classified

into six groups by one observer. The classification, at best, was vague and indefinite, but it represented the unbiased judgment of a single person. Inasmuch as the reproductions were classified and re-classified a great many times, small errors in the classification would be counterbalanced.

The first group contained fairly accurate reproductions of the original, diamonds of approximately the same size as the copy, having the sides and opposite angles nearly equal, and with a proper proportion between length and width. The second group contained figures inferior to those of the first group in size or symmetry, but representing a fairly high grade of ability. The reproductions that were less symmetrical than those of the second group were classified in the third and fourth groups. Figures showing some inequality between length and width were classified in the third group, while those of approximately unit proportion, square shaped figures, were classified in the fourth group. The reproductions placed in the fifth group were figures less symmetrical than those of the fourth group, and figures which had curved sides and rounded corners. The sixth group contained all figures which it would have been difficult to have recognized as intended for a diamond, figures having three, five or more sides, circles, ellipses, unfinished lines and eccentric figures.

The above classification did not offer an opportunity for a sharp grading between one group and another, but in general, the reproductions placed in the various groups from the first to the sixth, represented a decrease in the ability to copy the diamond. The justification of the method was not in the accuracy of the classification, but in the fact that the material was all classified by one observer (B), in such a way that he was in ignorance of the original rank that had been given the reproduction, of the experimenter who graded it, and of the character of the subject.

16% of the reproductions were classified in the first group, 21% in the second group, 20% in the third group, 17% in the fourth group, 9% in the fifth group, and 17% in the sixth group. (The irregularity of the distribution is due to the presence of the diamonds drawn by the Trenton subnormal group.)

After classifying all of the reproductions the ranks given to them by the different experimenters were then compared with the group in which they were classified. That the sliding scale classification used represented real differences between the reproductions is shown by the relative certainty of the experimenters' judgments. None of the reproductions classified in the first and second groups were ranked as failed by the four experimenters, while only one reproduction in the third group was ranked minus. 18% of the fourth group, 45% of the fifth group and 77% of the sixth group were ranked as failures. All of the sixth group diamonds that were ranked plus (23%), were so ranked by one experimenter, A.

To obtain a general estimate of the relative severity of the experimenters' criteria in making their judgments of passed or failed, the diamonds obtained by each experimenter from boys and girls were classified according to rank, plus or minus, and according to their group in the classification. From this it was possible to obtain an estimate of the passing mark of each experimenter. For example, the boys of experimenter B passed the test 72% of the time according to his ranking. Had B given credit for the first five groups and failed only the reproductions in the sixth, i. e., had his passing mark been the fifth group, 88% would have passed. Had his passing mark been the fourth group, 81% would have passed. If it had been the third group, 72% would have passed, while only 56% would have passed had it been the second group. Since 72% of B's subjects actually passed the test, his passing mark was the third group—in the long run, he would pass all diamonds in the first three groups and fail all in the last three. The differences between the experimenters on this basis are quite marked. The passing mark for C and D was the fourth group, while A's passing mark was the fifth group. B was the most severe, and A was the most lenient, with C and D between the two. The results were the same for both sexes.

Another test in which the influence of the personal equation might be looked for is that of copying designs from memory. The experimenter must here use his own judgment in marking

the designs passed or failed. Very little guidance is given by Binet's rule, which reads, "The test is considered passed when one of the designs is reproduced exactly, and half of the other is correctly drawn", or by the interpretation of this "half right" as applying "when two component parts are transposed or one component part omitted".

In order to test the experimenters' judgments in ranking this test, a scoring system was devised, which may be explained by reference to Figure 1, which gives the original copy and various duplicated portions. In scoring the reproductions of the pyramid section, 5 points were given when the reproduction of the asymmetry of the figures was nearly exact, as in no. 1, 4 points for a less perfect reproduction as in no. 2, and 3 points for a reproduction in which the rectangle fell in the center of the figure, as in no. 3. 1 point was deducted from this score for each failure to connect the corners of the rectangles as in no. 4 (which is modified from no. 3 and would therefore receive only 1 point), and no credit was allowed for "boxes" (no. 5), and other eccentric figures.

In scoring the more complicated design, 4 points were allowed for each of the "posts", ABCDE and JKLMN, or no. 6. 2 points were deducted for turning them in the wrong direction as in no. 7, (which is "post" ABCDE turned in the wrong direction), 2 points for failure to make the line AB penetrate DE as in no. 8, so that a combination of these errors, as in no. 9, would receive no credit, along with other eccentric reproductions as in nos. 10 A, B, C, D, E and F. 1 point was given for each of the lines EF and IJ, and 5 points for the "hump", FGHI. A continuous line from E to J as in no. 11 would therefore receive no credit, while a division of the lines, without the portion FGHI, as in nos. 12 or 13, would receive 2 points. An accurate reproduction of the portion EFGHIJ, as in no. 14, would receive full credit for all parts, 7 points, no credit being allowed for eccentric reproductions of the "hump" as in nos. 15 A, B, C and D.

The maximum credit for the test is 20 points, divided between the two figures on the proportion of 5 to 15, a fair proportion

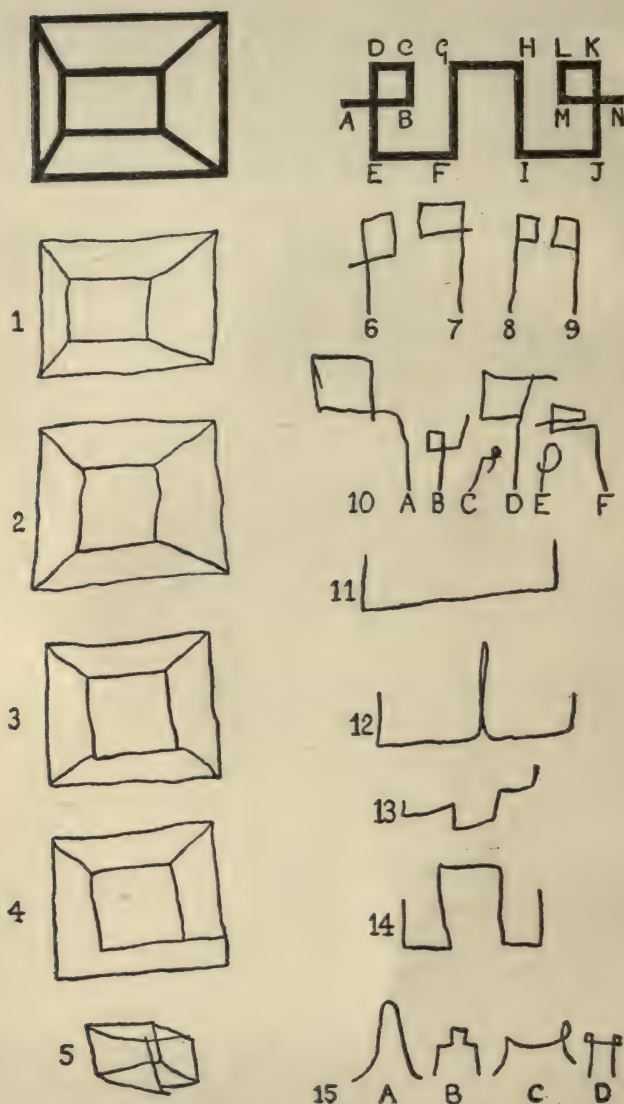


FIG. 1. *Method of Scoring Test of Copying Designs from Memory*

(in the writer's opinion) according to the relative difficulty of the parts. A design with "one component part omitted" would be scored 13 points according to this system, and one with "two

component parts transposed", 16 points, provided that the reproductions of the pyramid section were perfect in each case.

All the reproductions of the designs obtained from the Princeton and Trenton experimenting were then scored according to this system. The score of each subject of each experimenter in the Princeton series was then compared with the experimenter's ranking, which was recorded on the same sheet, and which was not seen at the time the designs were graded by the point system. From the number of times the test was given, and the number of times it was marked passed by the experimenter, the percentage passed was obtained for each experimenter for both sexes. The scores from all the designs from 0 to 20 were then classified according to the judgment passed or failed as given by each experimenter on subjects of both sexes. It was found that there were certain ranges where the experimenters' judgments coincided accurately, i. e. in the very low scores and in the very high scores. A certain range existed, approximately from 10 to 15 points, in which the same results would sometimes be ranked as passed and failed by the same experimenter at different times.

It was possible, however, to obtain a general estimate of the experimenters' criteria by a method similar to that used in the study of the diamond test. For example, B gave the test to boys 48 times, passing 40% of them. Had his passing mark been 18 (i. e. had he passed all subjects whose designs scored 18 points or better), 21% would have passed. Had his passing mark been 15 points, 35% would have passed. Had it been 13, 42% would have passed etc. B's passing mark would therefore fall between 13 and 15 points. In this way, by calculating the percentage passed at each score for each experimenter for both sexes, it was possible to obtain the passing mark of each group. The passing marks coincided very closely except in one case. With one exception the passing marks were around 12, 13, 14 or 15 points, for the boys and girls of all experimenters, i. e. the experimenters would, in the long run, rank all below this level minus and all above this level plus. The degree of correspondence was quite remarkable considering the fact that the

experimenters had very little on which to base their judgments.

The one exception is both striking and suggestive. C's passing mark for boys was 15 points, for girls 8 points. In order to receive a plus from C, boys would have to draw a much more accurate design than girls, or, in other words, a very faulty reproduction drawn by a girl would receive credit, while the same reproduction if drawn by a boy would invariably be failed. This deviation rests on a small number of cases. A gave the test to 24 boys and 21 girls, B to 48 boys and 33 girls, C to 28 boys and 22 girls, and D to 36 boys and 31 girls. A's results, although resting on a number of cases as small as C's, show no such deviation as those of the latter. On account of the small number of cases, this finding cannot be considered definite. It does, however, suggest the possibility of a difference in the experimenters' reaction to the sexes. An experimenter may show greater leniency to one sex than to the other, so that a supposed sex difference may be the results of an experimenter's reaction to the sex, rather than the sex's reaction to a test.

The test of using three words in a sentence ("Philadelphia, money and river") was given 279 times, and the sentences given by the subjects were recorded over half the time. Experimenter A gave the test 53 times, recording the result 36% of the time. B gave the test 95 times, recording the answer in 92% of the cases. C gave the test 56 times, recording the answer 23% of that number, and D gave the test 75 times, recording the response in 43% of the cases.

To obtain a check on the accuracy of the experimenters' scoring of this test, all of the recorded sentences were transcribed so that they could be studied and ranked without reference to the subject or the experimenter. The 162 recorded sentences were then marked plus or minus by one observer (B). This ranking was checked several times and then compared with the original ranking.

There was no disagreement between the judgments of the four experimenters and the one impartial observer in marking responses for the "ten year" credit. In marking for the "eleven year" credit, there were 8 disagreements out of the 162 judg-

ments, the 8 variations being evenly distributed among the experimenters. It may be concluded, then, that the influence of the personal equation is absent in this test, although there is ample opportunity for variation.

The detailed study of the foregoing tests has shown that the personal equation of the experimenters has a marked effect on the results of some of the tests. In the subsequent correlation of the tests with grade and sex the corrected score of these tests will be used. Only those definitions will be used which were recorded by the experimenters, and the ranking of the one observer will be followed. All reproductions of the diamond in the fifth and sixth group will be scored as failed, the others as passed. A reproduction of the designs scoring 15 or more points will be ranked as passed. The corrected results of the sentence test will be used.

To show that the effect of the personal equation of the experimenter is present or absent in the tests on which there is no actual record of the subject's response, is a more difficult problem. The most reliable method of showing the influence of this factor is that in which the reactions of similar groups of subjects, examined by different experimenters, are studied. The greater the similarity of the groups the more reliable the results. If two experimenters each examined 50 boys of 12 years of age from the sixth grade, their results should compare closely, and any difference could immediately be referred to a difference in the personal equation. However, if one examined boys from this grade and the other girls, the variations might be explained on the basis of sex differences. In the same way the results may vary with the age of the subject, and with his grade and nationality.

It is not possible in this study to obtain groups of a sufficient degree of similarity, in spite of the small number of children of non-English speaking parents, and the fact that the sexes may be treated separately. The subjects vary in age from 4 to 16, and in grade from the kindergarten to the sixth grade. A examined a very much younger run of subjects than B, C and D. The data of the four experimenters were treated by three meth-

ods, by comparing the per cent. that all boys and girls of each experimenter passed each test, by comparing the per cent. that selected subjects of each experimenter passed each test, and by comparing the per cent. that all subjects from 5 to 9 and from 10 to 13 passed each test. The sexes were separately treated in each method. None of the methods proved satisfactory, and it was found to be impossible to obtain an accurate quantitative estimate of the effect of the personal equation on each test. In certain of the tests, however, there were known differences of procedure which might have influenced the results, while the variations in the results of certain other tests were so striking that definite conclusions could be drawn.

One possible source of variation was the use of alternative questions in several of the tests. When an entire school system is examined, and the children learn that they will all be tested, the possibility is always present that they will inform each other of the nature of the tests, and the answers to some of the questions. The alternative questions were used to counteract the influence of this factor.

In the test of detecting absurdities in statements, ten or eleven statements were used, the experimenter choosing the five that he would give the subject. The statements varied greatly in difficulty and the experimenters did not use the same selection throughout the experiment. This test was given by B to 26 girls whose average age was 10.6 years, while D gave the test to 25 girls whose average age was 10.9 years. 65% of the girls examined by B passed the test, while only 36% of D's group passed. The variation between the experimenters might be due to the selection of absurdities of unequal difficulty, or to different criteria in grading the responses. The sources of variation are too large to admit of obtaining any reliable results from this test in correlating it with grade and sex.

75% of the girls to whom B gave the test of reconstructing dissected sentences passed, while only 28% of C's girls passed. The average age of the 26 girls to whom B gave the test was 10.8 years, and the average age of C's subjects 10.5 years. Part of the difference between these two experimenters is due to the

fact that more of B's subjects came from the fifth and sixth school grades. Some variation might have been due to different apparatus, B using cards with the sentences printed on two lines, while C had the sentences typewritten on one line. The sentences used by B were more legible, and, being broken into two lines, it was easier to grasp the individual parts as discrete units. Each experimenter used six sentences of varying difficulty so that some variation might be expected from the selection of the three sentences for the test. Whatever the cause of the discrepancy between the results of the two experimenters, it is obviously impossible to obtain any reliable conclusions concerning the correlation of this ability with age, grade or sex, on account of the presence of so many variable factors.

Three problems were used in the test of making change, 20c — 4c, 25c — 6c and 25c — 9c, the process of subtraction involved in each being of unequal difficulty. Certain variations occurred in the tests of comparing remembered objects and comprehending easy and difficult problem questions. Alternative questions were used in both of these tests, and variations might occur due to the relative severity of the experimenters' judgments in marking the responses passed or failed. None of the tests in which alternative questions were used will be treated in the subsequent discussion of the results.

At the close of the experiment, it appeared that a difference of procedure had existed between A and B in the test of indicating omissions in pictures. A and B both showed the three faces first, and the figure with the arms missing last, according to the standard procedure, but A, if his subjects failed to detect the parts omitted from the faces, would give them another trial after they had detected the missing arms. A gave this test to 51 boys and 33 girls, B to 30 boys and 30 girls, his subjects averaging about a year and a half above those of A. The test was passed by 76% of A's boys and 97% of A's girls, but by only 60% of B's boys and 63% of B's girls, showing that the difference of procedure had a most striking effect on the results. It is interesting to note what the effect of a difference of this magnitude would mean if the material from this test were used

as a basis of assigning it to the proper "age group" in the scale. If a test is to be considered normal for a given age if it is passed by 75% of the non-selected school children of that age, the test of indicating omissions in pictures would be a "six year" test for A, and an "eight year" test for B. The data from this test will not be treated in the subsequent discussion.

In the analysis of the results of the definitions test, it was found that certain differences existed between A, B and C in scoring the responses of the subjects as superior to use. No estimates could be made concerning D, for he did not record the actual responses. B, C and D gave this test to approximately the same range of subjects, averaging about 9 years. The corrected results of B and C show, in all, 28% of their subjects giving definitions superior to use, while 65% of D's subjects pass this test. Obviously D was very much more lenient than B and C.

The influence of the personal equation may or may not be present in the remaining tests. In the opinion of the writer it is not present to any marked degree. The data of the four experimenters were treated in several ways, and in none of these was it possible to demonstrate this influence. The writer's opinion, however, is more or less certain according to the test. The tests of repeating digits might show a slight difference between C's results and those of the other experimenters, a difference which could be explained by reference to the rate at which the digits were spoken. The results of experimenter D are slightly lower than those of the other experimenters in the tests of naming 60 words in three minutes and naming rhymes. Whether these differences are real or not, the writer does not know. The data from these tests are included in the subsequent study.

In the subsequent treatment of the results in terms of grade and sex, the material from the following tests will be treated.

- VI-2 and IX-2, Defining in terms of use and in terms superior to use.
- VII-1, Counting 13 pennies.
- VII-2, Describing pictures.
- VII-4, Copying diamond.
- VII-5, Naming four colors.

- VIII-2, Counting backward from 20 to 0.
- VIII-3, Enumerating the days of the week.
- VIII-4, Counting stamps (three singles and two doubles).
- VIII-5, X-3 and XII-1, Repeating 5, 6 and 7 digits.
- IX-3, Naming the day and date.
- IX-4, Enumerating the months.
- X-1, Naming the pieces of money.
- X-2, Drawing designs from memory.
- X-5 and XI-2, Constructing a sentence, containing one or two ideas from three given words.
- XI-3, Giving 60 words in three minutes.
- XI-4, Giving rhymes with "day", "mill" and "spring".

The treatment of the results of the definitions test will be confined to the recorded and corrected definitions of A, B and C. The results from the diamond test are based on the scoring system outlined, the passing mark being the fourth group unless otherwise indicated. The arbitrary point system of scoring the design test is used in the subsequent calculations, the passing mark, unless otherwise noted, being 15 points. The corrected scoring of the sentence tests will be used.

The foregoing study of the effect of the personal equation shows conclusively that in certain tests this influence is present to a very marked degree. The errors involved may be traced to three sources, to the apparatus used, to the technique of the experimenters in giving the tests, and to the experimenter's observation in marking the test passed or failed.

The error due to apparatus may result from a variation in the material itself, or from the calibration of different sorts of material as equal in difficulty, e. g.—alternative questions. The variation in the material used by B and C in the test of reconstructing dissected sentences illustrates the error due to defect in the material. The writer has seen apparatus for the line suggestion test in use, in which the last three pairs of lines were actually unequal, the difference between the pairs being above the threshold of discrimination. The subject with good discrimination will invariably fail this test when this faulty apparatus is used.

The error due to the use of alternative questions is more

common and therefore has more practical significance than defects in the material itself. There is a strong temptation for an experimenter, who believes a certain question to be unfair, to substitute another which seems to him to be of the same difficulty. In the study of the Trenton results, which will follow, it will be shown that the different questions included under the same test are not of the same difficulty. The question, "What would you do if you were delayed in going to school?" was passed by practically none of the normal children of 12, 13 and 14. If this question is changed to Goddard's (28) interpretation, "What ought one to do if he is afraid he'll be late for school?", the test is easily within reach of the 12 year children. The difficulty in the first test is caused by the word "delayed". Changing the structure of the test changes its nature completely. In this connection it is to be regretted that Town (72) in the appendix of her translation of Binet's 1911 scale, has changed the wording of some of the tests from that in the actual body of the translation. For example, the question "What would you do before taking part in an important affair?" (page 47) is changed to "Before taking part in something very important, what would you do?" (page 78), and "Why is a bad action done when one is angry, more excusable than the same action when one is not angry?" (page 47), becomes "Why do we more easily pardon a bad act done in anger than a bad act done without anger?", (page 79). The meaning is the same but the wording different, and in many cases success or failure in a test depends on the interpretation of a single word. If an experimenter using Town's translation were allowed to select his questions from the actual translation or the appendix indiscriminately, variations would, in all probability, result. The general proposition that there is no such thing as an alternative question, i. e. a question involving the same mental processes and having the same difficulty as another, could very easily be maintained. To avoid this error experimenters should adhere strictly to one wording and should never be allowed to substitute one question for another.

An example of the influence due to differences of the tech-

nique of the experimenters in giving the tests is afforded by the test of detecting omissions in pictures. This test is a "six year" test for A and an "eight year" test for B. Differences in procedure make it very difficult if not impossible to compare the results of one investigator with those of another. To eliminate this error, very careful and minute instructions should be published for the giving of each test. No edition of the Binet-Simon scale is entirely satisfactory in this particular.

Examples of errors due to the observation of the experimenters are afforded by the tests of copying a diamond and defining in terms superior to use. Errors due to observation may be avoided or minimized by increasing the number of grades of response with which the particular response in question may be compared. This principle is followed by Yerkes (82) in the arrangement of the Point Scale. In the diamond test, for example, Yerkes allows three grades of response while Binet allows but two—plus or minus. The accuracy of any measure increases with the number of gradations on the measuring scale, and the significance of the error of observation is diminished by decreasing the chances of wide displacement. In the tests in which a definite question is put to the subject, uniformity of scoring may be obtained by an accurate and painstaking cataloguing, and a subsequent classification and weighting of all the responses of a large number of subjects to each question. If the responses to a free association test may be classified into a relatively small number of groups, then the responses to a restricted association test could be classified into a much smaller number of groups. A sufficiently large number of responses will include practically all possible responses. In this way the chances of the error due to observation are diminished, while the adoption of a point system of scoring will minimize the effect of any errors that might be made.

The differences between the experimenters in this study are large enough to demonstrate the influence of the personal equation. Scientific procedure demands that the investigator who studies the results of the individual tests for the purpose of analysing the factors involved or for obtaining age norms should

demonstrate that the effect of the personal equation is not present in the results treated. The burden of proof should be on the person who maintains that the influence is not present. Negative results concerning the influence of the personal equation that are based on the method of comparing the total scores or "mental ages" of different experimenters should not be taken as conclusive, inasmuch as the experimenters may deviate in one direction in one test, and in the opposite direction in another, so that in a total score these deviations might equalize. In a study of this sort made on the basis of "mental ages," which has previously been reported, the writer (14) found no deviations between B. C and D, while deviations between these three experimenters do appear in the more detailed study of the individual tests. Studies of the individual tests can have no claim to reliability unless the personal equation has been eliminated.

The importance of the personal equation as a source of error in making diagnoses on the basis of the "mental age" of the subject is universally recognized by psychologists and almost universally ignored by medical men, field workers, school teachers and others who have had no experience in making mental measurements. Among psychologists there are two opinions concerning the solution of the difficulty arising from this source, the first, that of making certain allowances for the inexperienced examiners or establishing limits within which their opinions are valid, the second, that of removing the scale from their hands entirely.

Doll (22) in discussing criticisms of the Binet scale on the ground that diagnoses of normality and feeble-mindedness are made by inexperienced examiners urges "that those who are capable of doing good Binet testing of the mechanical sort without being clinical psychologists should report the findings of their examinations of children or groups in tables of related chronological and mental ages and not in terms of normality or abnormality. In their reports they can say with a high degree of certainty that those children who show an intellectual retardation of more than 3 years are feeble-minded, but they should not say that those who test less than 3 years retarded are backward or normal. In

the lesser degrees of retardation only the expert is capable of evaluating the details of a Binet test with any finality as to either diagnosis or prognosis." (page 607).

Doll also points out that Binet examiners who have worked in institutions give very reliable diagnosis, for they intuitively sense distinctions which inexpert laymen do not see. When the responsibility for the diagnosis is placed on the examiner in this way, the scale is treated as a qualitative instrument. This standpoint is quite different from that in which certain allowances are made for all inexpert examiners and the quantitative character of the scale preserved. Goddard (31) in a study of the personal equation based on re-testings of normal and feeble-minded individuals fixes the quantitative limits somewhat higher. "In all cases where a child tests four or more years behind his age, there is little danger of error in considering him feeble-minded, even though the test was made by a person who was not highly expert, provided such a person is able to use the test with reasonable intelligence. With the borderline cases, those who are two or three years backward, the best expert should be employed in the testing." (pages 76-77).

As early as 1910, before the scale had received very extensive application, Huey (35) took the stand that inexpert examiners should not use the scale. In discussing this point he said, "I would urge that these Binet tests must be used with judgment and trained intelligence, or they will certainly bring themselves and their authors into undeserved disrepute.—Results can be considered valid only when the tests are made by an experienced psychologist who has familiarized himself with Binet's directions, or by other competent persons who apply the tests under the direction and supervision of such a psychologist." (page 444).

Three years later, in referring to the reports that the medical inspectors in Pittsburgh were to take over the Binet testing in the schools, Whipple (78) says, "And we can only express our hopes that these reports are unfounded, or that at least those in authority may be led to understand that for a person, whoever he may be, without extensive psychological training to attempt to diagnose the precise mental status of a school child is about as

absurd as for a mere psychologist to attempt to diagnose incipient tuberculosis or any other obscure pathological condition." (page 302). The same position is taken by Whipple (77) in another editorial. "We have no quarrel with the use of the scale in the public school: properly used, it is of direct and practical value; but improperly used, it will become a farce which can but bring discredit upon psychology and retard the movement for its application to educational practise." (Page 119).

In defense of this position, Whipple calls attention to an error inherent in the procedure of all inexpert examiners. "There is nothing about the conduct of the Binet-Simon tests that is intrinsically difficult, yet there is a source of error inherent in the use of any psychological procedure, which, as experience shows, is surmountable only by drill in psychological experimentation. I refer to the difficulty of following directions. No one who has drilled students in the laboratory has failed to be struck with the impossibility of laying down fool-proof directions for the conduct by an amateur of a psychological test." (Page 119).

Kuhlmann (43) agrees with Whipple in this position. "The untrained examiner meets difficulties because he lacks the following: (a) Familiarity with the directions for giving the tests. (b) Familiarity with the rules for interpreting the responses of the children. (c) Ability to adapt the procedure in testing in special instances for which directions can not be given. (d) Ability to interpret responses in special instances for which rules can not be given. (e) Ability to adapt himself in attitude to the mental levels of children of different ages so as to obtain the best efforts from the child in each case. (f) General appreciation of the absolute necessity of adhering strictly to all rules of testing, and of careful, painstaking work. These deficiencies are of quite different degrees of importance. The last is, on the whole, the most serious and most frequent, and can be remedied only by extended laboratory training." (Pp. 255 and 256). In regard to the quantitative allowance that must be made for inexpert examiners, Kuhlmann's article affords the following, "The amount of error made by an examiner because of his lack of training seldom equals two years in the mental age; in the majority of cases it is less than one year." (Page 256).

IV. GRADE CORRELATIONS

The correlation between intelligence, as measured by the Binet scale, and school performance, as measured by age and grade standing, has been worked out by various investigators. In all cases intelligence was measured by the "mental age" or total score of the Binet tests, and pedagogical age by assuming that all children begin school at a certain age and should therefore be in certain grades at certain ages. Stern (62) has reviewed the work of Goddard (30), Binet (4), and Bobertag (10) in this field, with the general conclusion that the correlation is only moderately high. The number of children showing mental advance is in excess of those showing pedagogical advance, but very rarely do children showing pedagogical retardation show mental advance. The correlation is one-sided in that "inference from school performance to mental ability is safer than from mental ability to school performance." (Page 61). Stern accounts for the discrepancies on the ground that "performance in the school depends not only upon intelligence, but also upon certain other and quite different factors." (Page 63). These factors are strength of memory which plays a large part in school performance but correlates only to a moderate degree with intelligence, and other factors that have nothing to do with intellect but belong largely in the domain of the will—"the degree and duration of attention, industry and conscientiousness, sense of duty and capacity to fit into the social group." (Page 63). Stern concludes that "the lack of agreement between tests of intelligence and school performance is really calculated to increase our confidence in the psychological test-methods," (Page 64) that absolute correlation is not to be desired since that would mean that the tests were testing school performance only, and that the measure of intellectual ability was the school performance itself, the tests being superfluous.

More recently, Schmitt (57) has reviewed the work of God-

dard, Terman and Childs (66) and Dougherty (23) in correlating intelligence, as measured by the Binet scale, and school performance, and reaches conclusions quite opposite to those of Stern. The following quotations from Schmitt's monograph explain her view point. "Further doubt is cast upon the accuracy of the tests by the fact that judgments arrived at through their application do not coincide with that of the school concerning the same subjects." (Page 57). Concerning this lack of correlation Schmitt writes, "The Binet tests, therefore, while professing to test native ability are concerned very little with the education which all normal children have the native ability to acquire, and which is of much importance in civilized life." (Page 60). To the investigations cited Schmitt has added one of her own, in which the lack of correspondence between the Binet "mental age" and school grade is shown.

The writer is of the opinion that the method of correlating school performance with "mental age" fails to demonstrate either the adequacy of the Binet tests according to Stern, or the complete inadequacy of the tests according to Schmitt. For the demonstration of this point Schmitt's investigation may be discussed, inasmuch as it shows the most striking deviations between the measures of the two performances. Schmitt applied Binet's 1911 scale (Town's translations with modifications) to 150 children of superior social status. The following quotations indicate the status of the subjects. "The children who served as subjects for the tests comprised the Kindergarten and first six grades of a private school in Chicago." "They were the children of the professional class mainly. A few were children of successful business men who sought the best obtainable type of education for their children." (Page 2). The tests were applied at the close of an examination with the Healy-Fernald tests under rather unfavorable conditions as indicated by the following quotations,— "In the conduct of the two sets of tests the Binet-Simon tests were reserved for the last. By the time they were reached the child had been doing tests for an hour or more. In some cases there was too much restlessness and fatigue to

carry the child as far as the majority of his comrades in his grade were able to go and the tests were then discontinued." (Page 68 and 69).

The tests in the various age groups given to each grade were as follows,—Kindergarten, tests for V, VI, VII, VIII and IX years; Grade I, tests for V, VI, VII, VIII, IX, X and XII years; Grade II, tests for VI, VII, VIII, IX, X, and XII years. Grades III and IV, tests for VIII, IX, X and XII years; Grade V, tests for IX, X, XII and XV years; Grade VI, tests for XII and XV years. The "Adult" tests were also given to Grade VI as a class-room test.

Schmitt compared three measures, chronological age, school grade age and "mental age". The "mental age", in case a subject passed all tests in one group and failed one or more in a lower group, could be reckoned from two basal ages, these alternative rating being included by Schmitt. The summary of the results is as follows,— Comparing the Binet age to the chronological age, 14 (or 20)% are retarded, 26 (or 24)% are normal, and 58 (or 54)% are advanced. Comparing the school grade to the chronological age, (using 5 to 6.5 years as the normal age for the Kindergarten, 6.5 to 7.5 for Grade I etc.) 38% are retarded, 56% are normal and 4% are advanced. Comparing the Binet age to the school grade age, 2 (or 4)% are retarded, 25 (or 35)% are normal and 72 (or 60)% are advanced. The essential discrepancies are indicated by Schmitt by the following,— "Where the school grading shows 4% advanced over the normal for the chronological age, the Binet grading shows 58% over the chronological age and 72% over the age normal to the school grade." (Page 80.) The discrepancies thus indicated, although much larger than those of other investigators, agree with the general trend of results in that more children are shown to be advanced according to the Binet mental age than according to the school grade age. The results disagree with those of other investigators in finding a higher per cent. advanced by Binet age compared to chronological age.

The inadequacy of the methods employed in the investigations of Schmitt and others is seen when the measures are separately

studied. The use of the normal grade age as a measure of scholastic ability is false inasmuch as it rests on the assumption that all children enter school at a certain age, which is not the case. The measure of scholastic ability is the measure of the child's reaction to the subject matter of the grades, and that measure may be expressed only in the fact of promotion, non-promotion or (very rarely) double promotion, in other words, it may be expressed only in the relation of grade to the length of time in school. Furthermore, the two measures of scholastic ability, the age in grade method, and the grade progress method, are measures of an historically past performance not of present possibilities, and the true measure of an ability must indicate potential ability.

As measures of scholastic ability in terms of actual reaction, these measures present a distribution of general ability that is skewed toward the lower end, or in the direction of no ability. If a child enters school late, he presents a picture of retardation according to the age and grade method, while through any number of causes independent of intellectual ability, a child may present a retardation of at least a year according to either method. The possibilities for advancement are not as great, however, for advancement means forcing a child through a mass of subject matter, a process which the school is generally unwilling to undertake and the parent is generally unwilling to sanction. The school therefore presents a picture of ability in which promotion is normal, and non-promotion far more frequent than advance. If general ability is to be considered as distributed over any sort of a frequency surface, that surface will not take the form presented by the school measure in which the modal ability is almost completely the upper limit.

The measure of "mental age" has been shown to be one which varies from one chronological age to another in the form of its distribution. Normal children of 6 or 7 test over age, while those of 11 and 12 test under age. This abnormal distribution is due to two facts. In the first place, the tests in the younger years are too easy and those in the higher years are too difficult.

In the second place, the younger children have a wider range of tests beyond their average ability, so that exceptional subjects may display exceptional ability in a manner that is impossible if ability is measured by school progress, while older children have only a few tests within their range, the picture of advancement being excluded as in the measure of school ability. If the mental ages of a run of subjects of different chronological ages are combined, the frequency surface is normal, the error of the extremities balancing.

The investigators who have compared "mental age" with grade age, have compared two distributions, one of which is markedly skewed, the other normal, but false. The resulting finding of mental advance in excess of pedagogical advance has significance only insofar as it shows that a measure of general ability that will admit of exceptionally high performance is a better measure than one that precludes the possibility of such performance. The only significant finding is that pupils who show marked retardation in school rarely if ever show mental advance.

Applying the foregoing discussion to Schmitt's results in particular, all that has been said concerning the inadequacy of the age in grade method applies to her results. The age for entering school being 5, none of the subjects in the Kindergarten could be advanced, while those who entered late would be retarded. It is difficult to see how these young children would be able to make up their work in such a way as to show advance during the first two or three school years. The normal age for the sixth grade is from 11.5 to 12.5 years. Inasmuch as no grades were tested above VI, none of the 37 subjects from 11.5 to 14.5 could show an advance, and all of the 19 subjects from 12.5 to 14.5 would necessarily show retardation. Schmitt's results differ from those of other investigators in finding more subjects advanced according to Binet age in relation to chronological age. This deviation is probably due to the fact that she examined a superior selection of subjects, and to the fact that the XV year and "Adult" tests were used, so that the older subjects, who in general fall below their chronological age, had an opportunity to

better their scores. The discrepancy shown by Schmitt between school standing and the Binet tests does not demonstrate the inadequacy of the tests.

The final demonstration of a correlation between the Binet scale and school grade, rests not in comparing the total score or "mental age" with school grade, for that is susceptible to the errors of over-estimation and under-estimation according to varying chronological age, but in comparing the results of subjects in each grade on the individual tests. The tests may vary in their correlation with grade. Inasmuch as there is a general growth in age with grade, and a corresponding growth of intelligence with age, a test, in order to be an adequate test of intelligence, must show a correlation with grade. If the correlation is too high, however, the value of the individual test is in question for it would then be testing, not intelligence, but grade training. This criterion was actually used, though not stated, by Binet in his discussion of the results of Decroly and Degand (19), and in his revision of the 1908 scale, in which many of the tests that he considered to relate to school training were eliminated.

Studies of the individual tests in the light of school grade are not available. Decroly and Degand published in 1910 the results of an investigation on 45 children in a Brussels school, similar in character to that studied by Schmitt in Chicago. Binet discussed these results and those of other minor investigations in the Paris schools in considering the effect of environment on the results of the tests. Although he referred to school training as a factor, and classified the tests in which Decroly and Degand's subjects were superior, he gave no quantitative demonstration of the effect of this factor. The results of Decroly and Degand are based on too few subjects to admit of quantitative treatment. Chotzen (18) studied the tests by comparing the performance of feeble-minded individuals of the same mental age but of different chronological age. Although this method shows the effect of environment and maturity on feeble-minded individuals, it does not bear directly on the factor of school training. The foregoing

investigations will be discussed in this chapter only in their relation to the results of the particular tests. Schmitt, in her monograph, published tables showing the reaction of each subject in each grade to each test, the tables being discussed in the text. Although it was not Schmitt's purpose to determine the correlations between the various tests and grade, her data are available for a study of this sort, and the writer has taken the liberty of figuring them in this light, indicating at the same time Schmitt's interpretation of the grade factor, contained in the accompanying text. These data will be compared with the results of the Princeton investigation.

422 subjects of this investigation were distributed in the kindergarten, first six regular grades, minus grades and the special class of the Princeton Model School. 301 of the subjects (161 boys and 140 girls) were in the kindergarten and first six regular grades. The data obtained from the examination of these 301 subjects were classified according to the grade in which the subjects were found, and the percentage that the subjects of each grade passed each test was calculated.

Only those tests were studied which showed themselves to be free from the influence of the personal equation of the four experimenters. The elimination of the unrecorded results of the definitions test left a number of cases too small to be studied. To avoid the influence of the error due to incomplete data, the writer has calculated the percentage from only those tests that were given from 75% to 100% of the possible number of times. The data from the tests of repeating 5, 6 and 7 digits have been combined into one weighted measure. The procedure of the experimenters in giving these tests was to start within the subject's range and continue till he failed. If 5 digits were successfully repeated, 6 were given, and if these were passed, 7 were given. The results have been combined into one measure for the sake of simplicity, 1 point being allowed for the successful repetition of 5 digits, 2 points for 6 digits and 3 points for 7 digits, the weighting being roughly in accordance with the weighting in Goddard's scale, the tests being in the age groups

VIII, X and XII respectively. The measure of the ability of a group to repeat digits is the per cent. that the number of points scored is of the number of points possible (i.e. 6 times the number of subjects in the group).

The number of subjects in each grade (boys and girls shown separately) the average age of the subjects in each grade, together with the mean variation from the average are shown in Table 3.

TABLE 3

Number of Boys and Girls in Each Grade, and the Average Age of All Subjects in Each Grade.

Grade	Number of Boys	Number of Girls	Total No. of Subjects	Average Age	Mean Variation
Kindergarten....	20	12	32	5.64 years	0.46 years
Grade I.....	27	24	51	7.05 "	0.50 "
Grade II.....	16	24	40	8.16 "	0.65 "
Grade III.....	21	24	45	9.31 "	0.75 "
Grade IV.....	20	15	35	10.46 "	0.91 "
Grade V.....	24	25	49	11.71 "	0.99 "
Grade VI.....	33	16	49	12.81 "	1.06 "

The above table shows an increase of a year or more (actually from 1.10 years to 1.41 years) in the average age of the subjects in each grade. From this it is reasonable to expect that there is a general growth in intelligence correlating with this increase in age, or, in other words, to expect a correlation between the results of the individual tests and the grade in which the performance occurred. If the correlation is too high, it will indicate a dependence of that particular test on the subject matter of the grade. In Table 4 are shown the percentages that the subjects in each grade passed each test. The notes referred to in the margin contain the proportions passed for all other subjects for whom the percentages are not given, the percentages being given only for those groups to whom the tests were given from 75% to 100% of the possible number of times.

A study of Table 4 shows that the tests in general correlate with grade. The combined score of the test of repeating digits, for example, shows a growth from 6% to 78%, more rapid in the first three grades than in the last four. The tests vary in

TABLE 4

Percentage that Subjects in Each Grade Passed Each Test. 301 Subjects.
 Grades

Test	K	I	II	III	IV	V	VI	
VII-1, 13 pennies	72	96	100					Note 1
VII-2, Pictures	69	96	94					Note 2
VII-4, Diamond	46	75	88					Note 3
VII-5, Colors	72	90	97					Note 1
VIII-2, 20 to 0		9	53	80				Note 4
VIII-4, Stamps		13	50	78				Note 5
All digits, (combined)....	6	21	42	51	55	78	75	
VIII-3, Days of week....	16	45	90	100				Note 6
IX-3, Date		5	35	96	100			Note 7
IX-4, Months			28	84	90			Note 8
X-I, Money			20	36	57	82		Note 9
X-2, Designs				21	37	42	66	Note 10
X-5, Sentence (2 ideas)...				67	89	88	98	Note 11
XI-2, Sentence (1 idea)...				22	46	51	74	Note 12
XI-3, 60 words					63	63	87	Note 13
XI-4, Rhymes					67	63	76	Note 14

Note 1. Counting 13 pennies and naming colors given 20 times above II. Not failed.

Note 2. Describing pictures given 21 times above II. Not failed.

Note 3. Copying diamond given 25 times above II. Not failed.

Note 4. Counting from 20 to 0 given 18 times in K. Not passed. Given 31 times above III. Failed once.

Note 5. Counting stamps given 15 times in K. Not passed. Given 35 times above III. Failed 3 times.

Note 6. Naming days of week. Given 32 times above III. Not failed.

Note 7. Giving day and date given 5 times in K. Not passed. Given 56 times above IV. Not failed.

Note 8. Naming months. Given 26 times below II. Passed twice. Given 44 times above IV. Failed twice.

Note 9. Naming money. Given 26 times below II. Passed 3 times. Given 28 times in VI. Failed twice.

Note 10. Copying designs given 33 times below III. Passed 5 times.

Note 11. Sentence (2 ideas) given 32 times below III. Passed 12 times.

Note 12. Sentence (1 idea) given 32 times below III. Passed 4 times.

Note 13. Giving 60 words given 53 times below IV. Passed 19 times.

Note 14. Giving rhymes given 42 times below IV. Passed 26 times.

the number of grades taken to reach their maximum. The test of naming the day and date, for example, is failed by all subjects in the kindergarten, 95% of Grade I and 65% of Grade II, while only 4% of the subjects in Grade III and none of those in the

higher grades fail it. A sudden increase occurs between Grades II and III showing possibly the influence of grade training. The tests vary considerably in the degree of their correlation. An easily obtained measure of the degree of correlation is that of comparing the magnitude of the increases from grade to grade. For example, there is an increase of 61% (96%—35%) from Grade II to Grade III in the ability to pass the test of giving the day and date, and an increase of 16% (36—20%) between the same grades in the test of naming the pieces of money. The former test correlates higher with the influence of grade in this particular case than the latter.

In this manner the percentage difference between the performance of the subjects in each grade and that of the subjects in the preceding grade was obtained. All the increases or decreases in ability from one grade to another were thus obtained, these values serving as measures of the amount of correlation between the tests and the grades. 42 differences between the performance of the subjects in any grade and those of the next succeeding grade were thus obtained. In 4 cases there were actual decreases of 1, 2, 3 and 4% which were not significant. The difference ranged from —4% to +61%, the median being +19.5% ($Q=16.25\%$). Some of the differences between the grades might be due to the chance superiority of a particular grade. To overcome this chance variation, and to furnish another index of the growth of the various abilities, the differences were calculated by steps of two grades, i.e., subtracting the performance of the kindergarten from the second grade, the first from the third, etc. In this way, 26 differences were obtained varying from +9% to +91%, the median being +29% ($Q=18\%$).

Some of the differences noted are undoubtedly high enough to warrant the assumption of the effect of grade training on the tests. Just what tests show this effect is probably a matter of opinion. Allowance must be made for the growth of an ability independent of training. 25% of the highest increases from one grade to another were selected as being worthy of special consideration at least. A larger increase must be allowed be-

tween two grades. Those differences were considered worthy of special consideration that exceeded twice the value of the median of the one-grade differences or 39%. This manner of selecting the largest differences is quite arbitrary, but is justified by the outcome, for the tests that show the most significant increases according to this method show those increases in more than one step, so that the evidence is concentrated against a very few tests. In this way the significant values outweigh the less significant values and fair allowance is made for growth from one grade to another.

The following list includes the tests showing the greatest increases by one-grade and two-grade steps, together with the magnitude of the increases and the grades between which they occur.

One-grade steps. 25% of largest increases.	Two-grade steps. Increases greater than 39%.
+61% Date, II to III	+91% Date, I to III
+56% Months, II to III	+74% Days, K to II
+45% Days, I to II	+71% 20 to 0, I to III
+44% 20 to 0, I to II	+65% Stamps, I to III
+37% Stamps, I to II	+65% Date, II to IV
+30% Date, I to II	+62% Months, II to IV
+29% Diamond, K to I	+55% Days, I to III
+29% Days, K to I	+46% Money, III to IV
+28% Stamps, II to III	+42% Diamond, K to II
+27% 20 to 0, II to III	
+27% Pictures, K to I	

The above lists of increases are confined to but 8 tests. In all, there were 16 tests studied. According to the method of selecting the significant increases, 20 such values actually appeared. In this manner the evidence combines against a very few tests. Some tests appear in both lists and more than once in the same list. The most striking growth with grade is shown in the tests of giving the day and date, naming the months, nam-

ing the days of the week, counting from 20 to 0 and counting stamps. The tests of copying the diamond, describing pictures and naming money may or may not show this influence. The evidence is strongest in the case of the diamond test since that appears in both lists.

The foregoing method of selecting those tests which correlate with grade to such an extent as to indicate the influence of grade training is not conclusive, owing to the fact that there is also an increase in age from grade to grade. If a test showed a very rapid growth with age, and those ages fell for the most part in certain grades, then those grades would show an increase which might be wrongly assumed to be due to training. The tests of counting from 20 to 0 is a case in point. Yerkes (82) in Table 32, page 125, gives the percentage values for each test in the Point Scale, for English speaking boys and girls of each age. The test, of the twenty one tests included, that shows the most marked increase with age is that of counting backward, the values being as follows,— age 4=0%; age 5=3.5%; age 6=23.7%; age 7=45.7%; age 8=72.2%; age 9=96%; the values for ages above 9 being 97% or higher.

The age in grade distribution of the 301 subjects in this investigation is given in Table 5.

TABLE 5
Distribution of Subjects in Each Grade according to Chronological Age.
Grades

Age	K	I	II	III	IV	V	VI	Total
4	4							4
5	17							17
6	11	28	2					41
7		18	17	2	1			38
8		4	15	18	1			38
9			5	13	11			29
10		1		10	14	18	1	44
11			1	2	3	16	16	38
12					5	8	11	24
13						4	12	16
14						2	5	7
15						1	3	4
16							1	1
Total	32	51	40	45	35	49	49	301

The rapid growth of the ability in counting from 20 to 0, according to the method of comparing the subjects in each grade, was from 9% in Grade I to 80% in Grade III. From Table 5 it may be seen that practically all, (89%), of the chronological ages in Grades I, II and III were distributed in the ages 6, 7, 8 and 9, a chronological range coinciding with that in which Yerkes' results show the ability to develop. The growth of this ability might be due then either to age or to grade. For this reason, to arrive at any final conclusion, it is necessary to compare the subjects of the same age but in different grades. The treatment of the Princeton results according to this method follows, but the analysis of the data in this manner can have no great reliability owing to the small number of subjects in each group. The number of subjects in each group, (boys and girls shown separately), the average age and mean variation from this average are shown in Table 6.

TABLE 6

Number of Boys and Girls of Similar Ages in Different Grades, and the Average Age of the Subjects of Similar Ages in Each Grade.

Grade	Age	Number of Boys	Number of Girls	Total no. of Subjects	Average Age	Mean Variation
Kindergarten.	5	11	6	17	5.48	0.20
Kindergarten.	6	8	3	11	6.26	0.21
Grade I	6	14	14	28	6.59	0.17
Grade I	7	9	9	18	7.36	0.22
Grade II	7	7	10	17	7.56	0.24
Grade II	8	6	9	15	8.39	0.24
Grade III ...	8	8	10	18	8.60	0.22
Grade III ...	9	5	8	13	9.43	0.16
Grade IV	9	5	6	11	9.65	0.13
Grade IV	10	10	4	14	10.39	0.30
Grade V	10	7	11	18	10.54	0.25
Grade V	11	10	6	16	11.54	0.22
Grade VI ...	11	10	6	16	11.53	0.26
Grade VI	12	6	5	11	12.52	0.14

All chronological ages were computed in tenths of a year, so that a variation in age from 0.1 yr. to 0.9 yr. is possible within

TABLE 7

Actual percentage that each test was passed by the subjects of each age in each grade. 223 subjects.

AGE	Age 5	Age 6	Age 7	Age 8	Age 9	Age 10	Age 11	Age 12
GRADE	K	I	I	II	III	IV	V	VI
NO. OF SUBJECTS	17	11	18	17	15	18	13	11
Counting 13 pennies	71	91	96	94	100	100	Note 1	
Describing pictures	65	82	96	100	100	86	Note 1	
Copying diamond	50*	45	64	89	93	92	Note 2	
Naming colors	71	82	89	100	100	100	Note 1	
Counting from 20 to 0		16	0	63	43	83*	Note 3	
Counting stamps		13	12	29	60	85*	Note 4	
Repeating digits, (all)	6	6	21	19	42	46	47	51
Naming days of week	0	40	50	94	80	100*	Note 5	
Giving the day and date			29	20	20	88	100	100
Naming months			25	27	88	85	100	Note 6
Naming pieces of money				31	35	31	55	64
Copying designs				Note 10	89	62	91	43
3 words in sentence. 2 ideas				Note 11	33	8	45	50
3 words in sentence. 1 idea					Note 12		44	77
60 words in 3 minutes					Note 13		92	50
Giving rhymes							56	69

*The test of copying the diamond was given 59% of the possible number of times in K-5, counting from 20 to 0 and giving days of week, 66% in III-8, and counting stamps 72% in III-8. All other percentages are based on tests given from 75 to 100% of the possible number of times.

- Note 1. Tests of counting 13 pennies, describing pictures and naming colors each given 12 times above II-8. No failures.
- Note 2. Copying diamond given 15 times above II-8. No failures.
- Note 3. Counting from 20 to 0 given 16 times below I-6. Not passed. Given 31 times above III-8. Failed 4 times.
- Note 4. Counting stamps given 14 times below I-6. Not passed. Given 32 times above III-8. Failed 4 times.
- Note 5. Giving days of week given 32 times above III-8. No failures.
- Note 6. Giving date given 39 times below II-7. Passed twice. Given 36 times above IV-10. No failures.
- Note 7. Naming months given 24 times below II-7. Passed twice. Given 37 times above IV-9. Failed 4 times.
- Note 8. Naming pieces of money given 35 times below II-8. Passed 4 times. Given 14 times above V-11. Failed twice.
- Note 9. Copying designs given 26 times below III-8. Passed 5 times. Given 15 times above V-11. Failed 6 times.
- Note 10. Three words in sentence, 2 ideas, given 24 times below III-8. Passed 9 times.
- Note 11. Sentence, 1 idea, given same as 2. Passed 3 times.
- Note 12. 60 words in 3 minutes given 41 times below IV-9. Passed 10 times.
- Note 13. Giving rhymes given 37 times below IV-10. Passed 25 times.

each age group. That the subjects of the "same" age but in different grades are not exactly the same is shown in Table 6. The subjects of each age in the higher grades average from 0.01 yr. to 0.33 yr. different, with an average superiority of 0.19 yr. This difference, however, is about one fourth that between the subjects of different ages in the same grades, and may be called the same for practical purposes. For convenience, the groups will be referred to as K-5, II-7 etc., the first member referring to the grade, the second to the age. K-5 would mean the group of 5 year children in the kindergarten, II-7, the 7 year subjects in Grade II, etc. The actual per cent, that the subjects in each group passed each test was calculated and is shown in Table 7. Unless otherwise noted, the percentages are based on tests given 75% to 100% of the possible number of times.

Some of the groups from which results were obtained are too small to have great reliability, but the method is at least suggestive. The results of 14 groups are given. It is possible then to compare the results of subjects of 6 ages, (6, 7, 8, 9, 10 and 11), that are in different grades, and also to compare sub-

jects in all seven grades that are of different ages, and in this way to determine whether the dominating factor in the growth of any ability is that of grade or age. The reliability of the method rests only on its connection with that of the first method employed.

In answer to the question of whether the growth of ability in the test of counting from 20 to 0 is due to age or grade, a question which was unanswered by the first method, we may turn to the results shown in Table 7 in which the subjects of each age in each grade are shown. The test of counting from 20 to 0 was not passed by any of the 5 and 6 year subjects in the kindergarten. Comparing first the subjects of different ages in the same grade, the 7 year subjects in Grade I are 16% lower than the 6 year subjects in that grade, and the 8 year subjects in Grade II are 20% lower than the 7 year subjects in the same grade, the older subjects making a lower record in each case. Comparing the performance of the subjects of the same age but in different grades, the 7 year subjects in Grade II are 63% ahead of the subjects of the same age in Grade I, while the 8 year subjects are 40%¹ ahead of the subjects of the same age in Grade II. Allowing for the retrogression of the older subjects in each group, i.e. assuming that they should have done equally as well as the younger subjects in the same grade, the groups in Grades II and III are still 47% and 20% ahead of the subjects in the grades lower. The growth of ability in this test would therefore appear to be due to grade training.

A rapid growth of ability in the test of counting stamps occurred between Grades I and III ($37\% \text{ I-II} + 28\% \text{ II-III} = 65\% \text{ I-III}$), according to the first method, so that the same question arises as in the test of counting from 20 to 0. The test was not passed below group I-6. No growth with age is shown between

¹ This test was given to but 66% of the subjects in III-8, the experimenters assuming that the other 34% would pass. The score given, 85%, therefore represents the ability of the lowest selection of III-8 subjects, or the most conservative estimate of the ability of the whole group. The same applies to the other tests in III-8 given 66% and 72% of the time. In this way the hypothesis that the tests are not influenced by grade training is given the benefit of the doubt.

I-6 and I-7, but a growth of 31% appears between II-7 and II-8. A growth with grade of 17% is shown from I-7 to II-7 and of 25% from II-8 to III-8. This test shows therefore the operation of the two factors of age and grade training.

The improvement in ability in the tests of counting 13 pennies, describing pictures and naming colors, that was indicated between the kindergarten and Grade I by the first method, would refer to age rather than grade, for a greater increase in each test is indicated between K-5 and K-6 than between K-6 and I-6. Above I-6 these abilities are completely developed. It could be maintained that these tests are so completely within the ability of the groups that the effect of training would not be indicated. The test that is best adapted to show the influence of any factor on a group is one that is well within the ability of the group—the influence of the factor will be obscured if the measure is either too easy or too difficult. The test of copying the diamond is a case in point and one well worth study, for it has been attributed to the effect of training by various authors. All the reproductions of the diamond had been scored according to the arbitrary system outlined in the previous discussion of the personal equation. A control on the factor of difficulty was obtained by raising or lowering the passing mark in this test. The percentage passed was calculated for each group for each of the 5 possible passing marks. The relations indicated in Table 7, where the passing mark is Group IV, were not changed by this process of raising or lowering the passing mark. In all cases the influence of age was shown between groups I-6 and I-7, and the influence of grade shown between groups K-6 and I-6. The test was given to but 59% of the K-5 group, the experimenter assuming that the other 41% would fail, so that the percentages calculated represent the performance of the best selection of K-5 subjects, or, in other words, the benefit of the doubt is given to the hypothesis that the test is influenced by grade training. If the other members of K-5 had failed according to the experimenter's assumption, (and this assumption was quite justified for some had failed to draw the square), 29% of the group would have passed instead of 50%.

The influence of age indicated in this test is as great if not greater than that due to training.

The test of repeating digits, scored by the weighting system previously described, exhibits a slow but uniform progress throughout, the older subjects in each group making records that are about the same or slightly lower than those of the younger subjects in the same grade, an increase showing fairly regularly from grade to grade. The most marked increase in this ability appears between K-6 and I-6, and between I-7 and II-7, possibly indicating that the lack of familiarity with the use of digits in the lowest grades interferes with this test as a measure of auditory memory.

The test of naming the days of the week shows the most marked improvement with age (40%) from K-5 to K-6, practically no improvement (10%), from K-6 to I-6, no improvement from I-6 to I-7, a very marked increase with grade from I-7 to II-7, a drop from II-7 to II-8, group III-8 marking the complete development of the ability. The test would appear to be due to the combined effect of age and grade. The tests of giving the day and date and naming the months are passed only twice in the kindergarten and first grade, by about a quarter of the subjects in II-7 and II-8 without age increase, while the subjects in III-8 shows a most marked increase due to grade. Above III-8 these tests are seldom failed. The test of naming the pieces of money shows a slow growth from 8 to 11, the largest increases appearing from III-9 to IV-9 and from IV-10 to V-10, improvement with grade in each case. Copying the designs from memory shows a growth of 26% from 8 to 11, the development occurring in two age steps, from IV-9 to IV-10 and from V-10 to V-11.

The growth with age cannot be determined in the tests of constructing sentences from three given words, because they were given to too few cases below the third grade. The results do not show whether III-8 is exceptionally high or III-9 exceptionally low. Both tests show decreases in ability from III-8 to III-9 and from V-10 to V-11. The ability in the easier test is well within the range of the third and higher grades, showing, therefore, no

improvement. The improvement in the second test develops from 33% to 80% in three steps, correlating with Grades IV, V and VI in each case. The most vital question, that of determining whether or not the language training in the third grade helps to make the construction of a sentence possible, cannot be determined owing to the lack of material in the second grade. The experimenters' assumptions in not trying the test would indicate this fact, but this is not experiment. The same lack of material makes conclusions in regard to the rhyming test impossible. The performance of IV-10 is exceeded only by VI-11. The test of naming 60 words in three minutes shows two decided increases with age and one decided drop with grade.

The foregoing analysis is based on a number of subjects in each group too small to have any great significance. The general fact of the correlation of the tests with grade remains, and conclusions concerning what tests correlate too highly with training can be answered only by considering both methods of study, and by considering only the largest deviations. The two most striking instances are found in the tests of naming the months and giving the date. These tests undoubtedly relate almost entirely to training. Less striking but equally definite is the relation of the test of counting from 20 to 0 to training. The tests of naming the days of the week and counting stamps show the influence of age to an extent almost as marked as that of grade, so that while the development in these tests is rapid, the grade factor probably exerts only part of the influence. Conclusions concerning the other tests are largely a matter of opinion, and the opinion of the writer has been indicated in the detailed discussion.

A study of the tests in relation to grade by the first method employed may be made from Schmitt's results. The author gives, in Table I, II, III, IV, V, VI and VII on pages 70, 71, 73, 74, 75, 76 and 77 of her monograph, the results of each subject in each grade on each test. From these tables the present writer has calculated the percentage passed in each test. A study of this sort rests for its reliability on the accuracy of the published tables, and the facts indicated by the tables do not always coincide

with Schmitt's discussion.² The writer has followed the tables rather than the discussion in calculating the results. In the VIII-2 test where an alternative rank is given for counting from 10 to 0 instead of 20 to 0, the writer has considered success in counting from 10 to 0 as a failure in counting from 20 to 0. In the line suggestion test Schmitt recognizes two types of failure, the typical failure according to Binet of accepting the suggestion of the first three lines, and the failure due to the fact that the subject actually judges the lines unequal after studying them. The second type of response Schmitt marks as passed, using a special symbol. The writer has calculated these percentages separately, entering the first or Binet type of response under "Line suggestion A" in the table, and the second type under "B." The V year and Adult tests were omitted. All of the other tests were included that had been given over 70% of the possible number of times. Unless otherwise noted, each test was given 100% of the possible number of times. Table 8 shows the per cent. that Schmitt's subjects in each grade passed each test in Binet's 1911 scale (Town's translation with modifications). The table is given with the reservation that the tables from which the percentages were calculated might contain misprints, and that the writer's interpretation of the tables might be at fault.

Inasmuch as there are many differences in procedure in giving the tests, and in the character of the schools tested, the results of the two investigations are not comparable in respect to the percentage passed in one grade in one study with those in the same grade in the other study. The method used in determining the

² In the discussion (page 69) Schmitt gives 15 subjects in the kindergarten failing test VII-4. Table I shows 13. On the same page she gives 24 subjects failing VIII-4. Table I shows 22 failing. In discussing the results of Grade I (page 72) Schmitt states that there is "more than 50% of failure with the discrimination of weight", while Table II shows 35% failure. Again, the tests referred to specific school instruction by Schmitt are VII-4, VIII-4, and IX 1, 2, 3 and 4. On page 72, in discussing the results of Grade I, she says "the tests below ten years which depend upon specific instruction are usually not passed except the VII-4 test. The percentages passed are as follows: VII-4 = 85%; VIII-4 = 45%; IX-1 = 35%; IX-2 = 75%; IX-3=90%; IX-4=30%. "Usually not passed" includes, therefore, tests passed 75% and 90% of the time.

TABLE 8

Per cent. that Schmitt's Subjects of Each Grade Passed Each Test. 150 Subjects.

	Grades						
	K	I	II	III	IV	V	VI
Number of subjects	25	20	17	21	22	22	23
VI-1, Distinguishing morning, afternoon	96	100*					
2, Defining in terms of use	92	94*					
3, Copying diamond	76	94*					
4, Counting 13 pennies	92	100*					
5, Choosing prettier of faces	92	100*					
VII-1, Showing right hand	92	80	100				
2, Describing pictures	72	65	81				
3, Executing 3 commissions	92	95	100				
4, Counting stamps	48	85	100				
5, Naming colors	96	100	100				
VIII-1, Comparing remembered objects	92	100	100	100	100		
2, Counting backwards from 20 to 0	40	85	94	95	100		
3, Indicating omissions in pictures	100	95	94	100	100		
4, Giving day and date	12	45	94	100	100		
5, Repeating 5 digits	64	85	94	100	100		
IX-1, Making change	6*	35	71	95	86	100	
2, Defining in terms superior to use	39*	75	65	100	95	100	
3, Naming pieces of money	28*	90	94	100	100	100	
4, Naming the months	6*	30	71	95	95	95	
5, Comprehending easy questions	61*	100	100	95	100	100	
X-1, Arranging 5 weights		65	41	57	50	64	
2, Copying designs		10	35	57	45	32	
3, Detecting absurdities		60	88	100	100	100	
4, Comprehending difficult questions		85	100	100	100	100	
5, Constructing sentence. Two ideas		65	76	100	100	100	
XII-1, Resisting suggestion, A. (Binet scoring)	64*	76	52	41	14*	100	
B. Judgment error counted plus			100	86	100*		
2, Constructing sentence. One idea	57*	71	95	95	100*	100	
3, Giving 60 words in three minutes	43*	82	62	100	95*	96	
4, Defining abstract terms	7*	29	52	73	95*	100	
5, Reconstructing dissected sentences	0*	6	10	23	81*	78	
XV-1, Repeating 7 digits						62*	78
2, Rhyming words with "obey"						86*	70
3, Repeating a sentence of 26 syllables						10*	17
4, Interpreting pictures						14*	70
5, Solving problems from various facts						62*	70

Note.—All tests except those marked (*) were given all the possible number of times. The VI year tests were given 90% of the time in Grade I, the IX year tests 72% of the time in the kindergarten, the XII year tests 70% of the time in Grade I, and the XII and XV year tests 95% of the time in Grade V.

correlation of the tests with grade is the same as that used in the first method of treating the Princeton data, that of comparing the differences between grades by one-grade and two-grade steps, of selecting an arbitrary standard for detecting exceptional growth, and of comparing the resulting lists. The differences between the performance of each grade and the next succeeding grade were calculated. These differences, 100 in number, ranged from -24% to $+62\%$, the median being $+5\%$ ($Q=10.75\%$). The run of differences differs from that found in the Princeton study in two respects, in having a lower median and variability, and in containing more minus deviations. The lower median and variability is due to the fact that the tests were given over a wider range, the Princeton tests being given only on the "up slope" of the growth curve, or not being given when the tests were any distance above or below the probable range of ability of the group. The Princeton results showed only 4 minus deviations of 4, 3, 2, and 1% respectively, while Schmitt's results show 15 such deviations, 6 of them being 10% or over. These deviations are probably due to the smaller number of subjects, and if due to chance, should be counteracted by the precautionary measure of combining the indices of correlation into two-grade steps. 71 two-grade differences were obtained ranging from -25% to $+82\%$, the median being $+10\%$ ($Q=16.5\%$). 4 measures were still in the minus direction, one of these, -25% (Design III to V) is probably significant, the other values of -6% , -5% and -4% having no significance. Inasmuch as the variability of the series is lower, those differences were considered to be worthy of special study that had the value of $2Q+M$, or were in excess of the interquartile range plus the median. The lists of tests that appear as showing marked growth with grade according to the two methods are as follows:

One grade differences higher than $2Q+M$	Two grade differences higher than $2Q+M$
+62%, IX-3, Money, K to I	+82%, VIII-4, Date, K to II
+58%, XII-5, Dissected, IV to V	+71%, XII-5, Dissected, III to V
+49%, VIII-4, Date, I to II	+66%, IX-3, Money, K to II
+45%, VIII-2, 20 to 0, K to I	+65%, IX-4, Months, K to II
+41%, IX-4, Months, I to II	+65%, IX-4, Months, I to III
	+65%, IX-1, Change, K to II
+39%, IX-5, Comprehension, K to I	+60%, IX-1, Change, I to III
+39%, XII-3, 60 words, I to II	
+38%, XII-3, 60 words, III to IV	+55%, XII-5, Dissected, IV to VI
+37%, VII-4, Stamps, K to I	+55%, VIII-4, Date, I to III
+36%, IX-2, Definitions, K to I	+54%, VIII-2, 20 to 0, K to II
+36%, IX-1, Change, I to II	+52%, VII-4, Stamps, K to II
+35%, IX-2, Definitions, II to III	
+33%, VIII-4, Date, K to I	+47%, X-2, Design, I to III
	+45%, XII-4, Abstract Def., I to III
+29%, IX-1, Change, K-I	+44%, XII-4, Abstract Def., II to IV
+28%, X-3, Absurdities, I to II	+43%, XII-4, Abstract Def., III to V

A study of the above lists shows, as in the similar study of the Princeton data, that although the method of selecting the exceptional tests is an arbitrary one, the method is justified in practice, for only a few tests (13) appear in the lists as significant. In all, there were 34 tests³ studied, and 30 differences were considered large enough to be significant. These 30 differences were confined to 13 tests. The tests of naming 60 words and defining in terms of use drop out of the first list owing to the elimination of the errors of negative correlation. The design test is both positive and negative, the ability increasing from Grades I to III and decreasing after III. The test of defining abstract terms appears according to the second method because the ability increases with grade from 7% in I to 95% in V by

³ No differences were calculated from the line suggestion test owing to the possibility of misinterpreting the symbols. Schmitt notes the difference in the character of the responses from the suggestion error to the judgment error in passing from Grade II to III. The scoring of the suggestion error in the tables shows an inverse correlation with Grades II, III, IV and V, and a sudden change again from 14% in Grade V to 100% in Grade VI, so that there is probably a mistake. The scoring of the responses to this test according to the strict Binet ruling would make the "mental ages" lower, for many cases would then have basal X.

increases of approximately 25% in each grade. No conclusions may be drawn concerning the easy comprehension test and the absurdities test. The 20 remaining differences are confined to 7 tests, those of naming the day and date, naming the months, counting from 20 to 0, counting stamps, naming money, reconstructing dissected sentences, and making change. The first four were included in the five found to show the most marked influence of grade in the Princeton study. The test of naming the pieces of money did not show a marked relation to grade in the latter study, but this difference might be one of school curriculum. The test of naming the days of the week is not included in Binet's 1911 scale.

In the Princeton study alternatives were used in the making change question so that no data from this test were included in the quantitative study. These data show the ability in this test developing in the second and third grades, the test being passed only twice in the kindergarten and first grades, and generally passed above the third. The data in the test of reconstructing dissected sentences show very few passing the test below grade V with approximately three fourths passing in V and VI. In so far as the Trenton experimenting was applied to a few subjects in the regular grades below the seventh, this test was rarely passed in the third and fourth grade, passed about 5% in V, and almost universally passed in VI, VII and VIII. The number of subjects in each grade is small in the Trenton experiment, but each test was separately scored, i.e. each part of the dissected sentence test, each part of the absurdity test etc. Each of the three parts of the dissected sentence test showed the same growth between the same grades, and this growth was more marked than that in any other test. The evidence concerning these two tests, therefore, supports the evidence from Schmitt's results.

The quantitative analysis of the Princeton data and Schmitt's data would indicate that the tests of counting stamps, counting from 20 to 0, naming the days of the week, giving the day and date, naming the months, naming the pieces of money, making change and reconstructing dissected sentences were influenced to a considerable extent by grade training. The performance in

certain of these tests (days, date and months) may be the result of specific school training in the tests themselves, while others (perhaps the tests of counting stamps, counting from 20 to 0, and reconstructing dissected sentences) may involve a transfer effect in the application of the content of the grade in a new way. The fact that the tests correlate very highly with grade training does not show that the tests are worthless, but it does show that they should, perhaps, be placed in another scale, or should at least be placed on a different footing than those that test capacity irrespective of attainments.

One of the best tests⁴ of intelligence is the determination of what an individual can do with the training he has received, but tests of this sort rest on the assumption that the individual's opportunities have been determined. The importance of tests of information in cases of alienation presenting a picture of deterioration is recognized. The important change to be made is not the elimination of such tests from intelligence scales, but their standardization on a different basis. The diagnostic value of such tests rests not in the mechanical memorizing of a time series such as that of the months, but in the ability to apply such a series. In pointing out this fact Katzenellenbogen (37) suggests that the months test be given in some such manner as "If somebody asks you in November to return three months later, what month would it be?" Decroly and Degand also suggest that the mechanical tests of counting and naming the days of the week and months be modified in some such manner.

⁴ The writer recalls two cases in which the failure in tests which involved the application of training was very significant. The first was that of a woman of about 30, a parole patient in a hospital for the insane, who had never shown any marked symptoms other than a history of intellectual inferiority. This patient passed practically all of the Binet tests in the IX, X and XII year groups, but failed completely in the test of making change. This observation was later checked up. Another case of a woman of 22, in the same hospital, presented a border-line psychoneurotic picture perhaps, but no marked symptoms other than a history of intellectual inferiority. She passed in a great many of the difficult tests in the upper years but had great difficulty in telling time. Both cases had lived under very good home conditions and had mingled with people of ability. A great many tests of capacity were given, but the most illuminating evidence of their mental status came from the two tests mentioned.

Comparing the conclusions of this study with other investigations, the agreement is fairly close. Schmitt's results do not support her suggestion that the definitions test relates to specific school instruction. The other tests which she refers to this factor (stamps, date, 20 to 0, change, months and money) show the influence to a marked extent. Binet in classifying some of the tests referred the tests of copying a sentence, reading for memories, writing from dictation, copying a diamond, counting backwards and making change to scholastic training. The first three tests were not included in this investigation. The diamond test showed the influence of age to be as great if not greater than that of school training. The last two tests showed a marked influence of training. Binet referred the tests of counting 13 pennies, naming four colors, naming the days of the week and enumerating the months to home training. The last two showed a marked influence of school training. The results of the present investigation agree with those of Chotzen in finding no effect or very little effect of training in the tests of copying the diamond, repeating digits, describing pictures, counting 13 pennies, naming colors, comparing remembered objects, defining in terms of use and superior to use, and in finding marked influence of this factor in the test of naming the days of the week.

The methods used in analysing the results, especially the second method, reveal several suggestive relations between the tests and the school grades. There is a general correlation between the tests and the grades, a correlation that is very necessary to establish, for there is also a general correlation between intelligence and grade. In analysing the results of the individual tests by comparing the results of subjects of the same age in different grades, and of subjects of different ages in the same grade (Table 7), it was seen that, as a general rule, the growth in any particular ability occurred in passing from grade to grade, not in passing from age to age within one grade. In fact in only half of the cases in which the subjects of two ages in one grade may be compared do the older subjects make records that are higher than those of the younger ones, and only 10% of these gains are over 20%. If the groups were considered to be equal in all

cases in which their records were within 10% of each other, equality occurs in exactly 50% of the cases. Of the remainder, 20% of the groups were lower, while in only 30% of the cases are the older subjects actually higher than the younger subjects of the same grade. Some of the cases of retrogression could well be accidental, but they occur too frequently to be due entirely to chance.

Applying the same general method to the cases in which groups of the same age but in different grades were compared, 5% of the groups in a higher grade showed lower scores, the results correspond in 43% of the cases, while 52% showed definite improvement. This might indicate that there is a higher correlation between the tests and grade than between the tests and age. The fact that the comparison of children of different ages in the same grade showed the older children making lower records in 20% of the cases, equal records in 50% of the cases and higher records in only 30%, would confirm the general diagnostic value of the tests if Bonser's interpretation of this phenomena is correct. Bonser (12) applied various sorts of reasoning tests to children in the fourth, fifth and sixth school grades. In summarizing the results of the tests in the different grades, he says, "In the contrast with grade progress and progress with age, in the generally superior showing made by the younger groups of children of any grade when contrasted with the older pupils of the grade, and in the fairly substantial percentage of pupils from lower grades found in the highest quartile of ability for all, it is shown that native capacity is measured to a high degree by the tests."

In conclusion, the results shown in this chapter would indicate a correlation between the individual tests studied and the school grades, this correlation being high enough in some cases to show the actual effect of training. In answer to the general objection that since one demonstration of the accuracy of the tests rests on their correlation with school grades, the school grades are the real measure of intelligence and the mental tests superfluous, it is only necessary to point out that intelligence tests, besides affording the opportunity for accurate standardization,

also detect the subject's potential abilities independent of his past performance. The school measure indicates mental defect in cases of gross retardation, but it does not indicate exceptional ability.

Schmitt's contention that the school represents a standard environmental situation, and a measure of a subject's ability should include a measure of the adequacy of his reaction to this situation, is well founded. It is not, however, a criticism of the Binet scale, for the scale aims to test native capacity. At the Buffalo conference (15) on the Binet scale, the following question was raised,—“What is it, after all, that the scale aims to test?” The question was answered by “We believe that current misconceptions as to the aim of the scale should be removed. It is not intended to test the emotional or volitional nature, but primarily intelligence (judgment).” To this list might be added the assertion that the scale was not intended to test a child's reaction to the school situation, or to furnish an outline for taking a record of his life history.

Rogers and McIntyre (54) would also have mental tests include tests dependent on both school and home training. This general trend of present day discussion is a reversion to Binet's 1908 type of scale, a tendency to which Binet was in opposition. The probable solution rests in eliminating from the scale the tests involving training, and in constructing a standardized scale of another sort for the estimation of the individual's reaction to the school situation in terms of the length of time that he has met that situation. That such a scale is not a matter of speculation is shown by the number of scales now on the market for measuring handwriting, spelling, composition, arithmetical ability, etc. Tests of native capacity and tests dependent on school and environmental training cannot be standardized on the same basis, for they are essentially different measures. Measures of the first sort may perhaps be correlated with age, while measures of the other sort can be correlated only with opportunity.

V. SEX DIFFERENCES

The investigators who have studied the influence of sex differences on the Binet-Simon tests have used two methods, that of comparing the "mental ages" or total scores of subjects of each sex, and that of comparing the per cent. that the subjects of each sex pass each test. The first method throws no light on the individual tests, inasmuch as one sex may be superior in one test and inferior in another so that the total score will balance the influence of this factor. Inasmuch as the scale is founded on the principle that sex differences do not exist, it is important to study the individual tests, and to determine the accuracy of this assumption.

The Princeton data are available for a study of this sort. 352 subjects (187 boys and 165 girls) between the ages 6 and 12 were examined. The method of study adopted was that of comparing the results of non-selected boys and girls of each age, and, as a check on this method, of comparing the results of selected boys and girls of four ages.

Inasmuch as the subjects of each chronological age are distributed over a range of one year (the 6 year subjects for example being distributed from 6.0 to 6.9), the actual average age of the subjects of each age was computed to make sure that no differences might appear due to the chance selection of subjects at either extreme. These averages are shown in Table 9.

TABLE 9
Actual Average Chronological Age of Boys and Girls in Each Age Group.

	BOYS		GIRLS	
	Number of Subjects	Average Age (M. V.)	Number of Subjects	Average Age (M. V.)
Age 6	37	6.58 (0.20)	23	6.51 (0.20)
Age 7	29	7.50 (0.29)	31	7.39 (0.26)
Age 8	24	8.48 (0.29)	28	8.48 (0.22)
Age 9	20	9.46 (0.27)	22	9.54 (0.26)
Age 10	31	10.46 (0.25)	23	10.37 (0.30)
Age 11	28	11.59 (0.22)	20	11.52 (0.27)
Age 12	18	12.43 (0.30)	18	12.57 (0.24)

A perusal of this table shows that the subjects agree closely both in their average and in their variability. The 12 year boys are actually 0.14 yr. younger than the girls of the same age group. The 7 year boys are 0.11 yr. older than the 7 year girls. All other differences are less than 0.10 yr. The correspondence is close enough for all practical purposes, but these differences must be taken into consideration before drawing final conclusions.

The 352 non-selected subjects from 6 to 12 were distributed throughout the kindergarten, special class, and first six minus and plus grades as shown in Table 10.

TABLE 10
Age in Grade Distribution of 187 Boys and 165 Girls, 6 to 12 Years of Age.

Age	6		7		8		9		10		11		12		Totals
Sex	B	G	B	G	B	G	B	G	B	G	B	G	B	G	
Special Class	2		1		3		3				3		1		13
Kindergarten	8	3													11
Grade I-	13	4	8	9	1	1	1								37
Grade I	14	14	9	9	3	1			1						51
Grade II-			2	2	2	4		2							12
Grade II		2	7	10	6	9	2	3			1				40
Grade III-					3		4	2	1	1		1			12
Grade III			2		8	10	5	8	5	5	1	1			45
Grade IV-							1		5	1	2	2	3	1	15
Grade IV			1		1		5	6	10	4	1	2	3	2	35
Grade V-									1	1		1	3	3	9
Grade V									7	11	10	6	2	6	42
Grade VI-											1		1		2
Grade VI									1		10	6	6	5	28
Totals	37	23	29	31	24	28	20	22	31	23	28	20	18	18	352

It is generally conceded that a difference exists in the reactions of the sexes to the school curriculum, the girls in the long run making better progress in school work than the boys. A study of Table 10 shows that in general the girls have a slightly higher distribution than the boys, these relations being more clearly indicated in Table 11 in which the average grade of the subjects of each age and sex is shown. In computing the average grade, the kindergarten was counted 0; Grade I—, 0.5; Grade I+, 1.0; Grade II—, 1.5; etc. Each subject in the special class was assigned a grade 0.5 lower than the lowest subject of his age (0 being the smallest value given), on the theory that each subject in

the special class was less satisfactory than any of his comrades in the regular class. The fact that there were no girls in the special class would cause an unduly exaggerated difference between the average grades of the boys and girls. For this reason, the average grades of the boys, including and excluding the special class cases, were separately figured, these values being separately shown in Table 11 under Boys A (the average grade including the special class cases), and Boys B (the average grade excluding the classes). Had the special class subjects been in the regular grades, they would have lowered the average of each group, so that the two values may be taken only as limits, the values under "Boys A" being the lower limit, and those under "Boys B," the upper limit.

TABLE 11
Actual Average Grade of Boys and Girls in Each Age Group.

	BOYS A		BOYS B		GIRLS	
	No.	Average Age (M. V.)	No.	Average Age (M. V.)	No.	Average Age (M. V.)
Age 6	37	0.55 (0.34)	35	0.59 (0.33)	23	0.87 (0.35)
Age 7	29	1.24 (0.64)	28	1.29 (0.63)	31	1.31 (0.65)
Age 8	24	1.94 (0.91)	21	2.21 (0.65)	28	2.25 (0.59)
Age 9	20	2.48 (1.04)	17	2.91 (0.69)	22	2.98 (0.62)
Age 10	31	3.92 (0.71)	31	3.92 (0.71)	23	4.19 (0.80)
Age 11	28	4.66 (1.20)	25	5.04 (0.77)	20	4.83 (0.88)
Age 12	18	4.72 (0.91)	17	4.82 (0.88)	18	5.03 (0.59)

Table 11 shows that the scholastic ability of the girls as indicated by the average grade is uniformly higher than that indicated by the lower limit of the boys, and is below the upper limit of the boys in only one case (at 11 years). A slight sex difference in school work in favor of the girls may therefore be assumed at the outset. It is significant that the upper limit of the 11 year boys is higher than that of the 12 year boys, and that the lower limits show a difference of but 0.06. This would indicate a poor selection of 12 year boys, or a superior selection of 11 year boys. Both measures of the scholastic ability of the boys show a generally higher variability than that of the girls.

From Table 9 it may be seen that the growth in the actual average age of each sex is not uniform from year to year, the minimum increase for boys being 0.84 yr. (from 11 to 12), and

for girls 0.83 yr. (from 9 to 10), while the maximum increase for boys is 1.13 yr. (from 10 to 11), and for girls 1.15 yr. (from 10 to 11). A more marked lack of regularity in the growth of scholastic ability from year to year as measured by the average grade is shown in Table 11, no increase being shown by the 12 year boys over the 11 year boys, while the 10 year boys show an increase of 1.44 to 1.01 grades over the 9 year boys. In the same way the 10 year girls show an increase over the 9 year girls that is nearly three times that of the 7 year girls over the 6 year girls, while the increase of the 7 year girls over the 6 year girls is twice that of the 12 year girls over the 11 year girls. These relations indicate that the selection of subjects is not uniform at each age. The subjects of any one age may be either a superior or inferior selection of all children of that age, and there is no reason for supposing that this random sample of superior or inferior subjects of any age will correspond to a similar sampling of the subjects of the opposite sex of the same age.

The process of calculating the percentage that the boys and girls of each age pass each test is extremely simple, but the conclusion, that the differences found between the percentage passed by the sexes at each age may be attributed to sex differences, is not justified unless all the variable factors are known.

A previous chapter showed variations in the tests due to the influence of the personal equation of the experimenters. To avoid this variable influence, only those tests were studied that showed that they were free from the influence of this factor. Inasmuch as each experimenter examined approximately the same number of boys and girls of each age, any influence of this factor would be equalized, provided, of course, that there were no differences in the reaction of the experimenters to the two sexes. In the detailed study of the design test, it was found that experimenter C was more lenient in marking girls than boys. The possibility of a similar interpretation in a few other tests was suggested, but not demonstrated. In analysing the results for sex differences, however, the possibility of such an interpretation must be kept in mind.

Another possible source of error is that due to incomplete data.

The experimenters, in giving the tests, would give only those within the approximate range of the subject, so that each test would be given to a superior selection of children below the normal range of the test, and to an inferior selection of subjects above this range, a process tending to make the apparent growth of an ability less than the probable real growth. In comparing the results of the sexes, however, it is not necessary to have accurate results on the growth of an ability, but results which have the same determining factors. If the experimenters gave the test to approximately the same proportions of boys and girls at each age, a comparison of the percentage passed is legitimate, even if a small proportion of the whole group were actually tested, for the proportion would include the same selection of subjects. The number of boys and girls at each age, and the percentage that each test was given to these subjects are shown in Table 12. The test of counting 13 pennies, for example, was given 37 times to 6 year boys, or 100% of the possible number of times, while the test of counting from 20 to 0 was given 27 times to the same group, or 73% of the possible number of times. Column A shows the total number of times each test was given to all of the boys and girls. Column B gives the average age of all the boys and girls to whom each test was given. The average given in this case is not the actual average derived from the actual chronological age of each subject figured in tenths, but the weighted¹ average, the whole numbers 6, 7, 8, 9, 10, 11, and 12 being used.

Table 12 shows a very close correspondence between the percentage that each test was given to boys and girls of each age, so that the error due to incomplete data, though present, is present to the same extent in the results of both sexes, and may be disregarded. A fairly close correspondence in the average age of all the boys and girls to whom each test was given is also indicated in Table 12. In the test of counting stamps there is an

¹ For example, in the test of counting 13 pennies, the average age of the boys to whom the test was given is,—

$$\frac{(37 \times 6) + (28 \times 7) + (16 \times 8) + (8 \times 9) + (7 \times 10) + (3 \times 11) + (1 \times 12)}{100} = 7.33 \text{ years}$$

TABLE 12

Percentage that Each Test Was Given to Boys and Girls of Each Age, the Total Number of Times Each Test Was Given to Each Sex and the Average Age of All Subjects of Each Sex to Whom Each Test Was Given.

Chronological age		6	7	8	9	10	11	12	A Total number of times given	B Average age of subjects. (weighted)
Number of subjects	Boys	37	29	24	20	31	28	18	100	
Number of subjects	Girls	23	31	28	22	23	20	18	90	
Counting 13 pennies	Boys	100	97	67	40	23	11	6	100	7.33
	Girls	100	94	68	41	26	10	11	90	7.56
Describing pictures	Boys	100	90	67	45	26	11	6	100	7.38
	Girls	100	94	68	41	30	20	11	93	7.66
Copying diamond	Boys	100	93	63	60	32	14	17	108	7.30
	Girls	100	94	61	64	35	20	11	97	7.74
Naming colors	Boys	100	93	67	45	23	11	6	100	7.35
	Girls	100	94	68	41	30	15	11	92	7.62
Counting from 20 to 0	Boys	73	97	83	80	61	21	44	124	8.18
	Girls	74	71	79	77	52	35	28	102	8.25
Counting stamps	Boys	65	97	88	80	61	21	44	122	8.23
	Girls	83	87	79	82	52	35	39	112	8.23
Repeating all digits	Boys	95	100	100	100	100	100	100	185	8.75
	Girls	96	97	96	100	100	100	100	162	8.78
Naming days of week	Boys	92	100	83	80	61	21	44	132	8.04
	Girls	96	90	82	82	52	35	28	115	8.10
Giving day and date	Boys	43	76	88	95	84	64	89	138	8.10
	Girls	78	77	93	100	78	75	72	136	8.70
Naming the months	Boys	41	79	79	95	81	54	78	130	8.90
	Girls	39	65	93	100	70	65	61	117	8.84
Naming money	Boys	27	62	67	90	97	64	78	124	9.21
	Girls	43	39	86	95	100	80	67	118	9.11
Copying designs	Boys	16	31	67	85	94	79	78	113	9.56
	Girls	26	19	57	86	96	80	67	97	9.46
3 words in sentence	Boys	8	31	63	90	100	93	100	120	9.79
	Girls	26	26	68	86	100	95	94	111	9.36
60 words in 3 minutes	Boys	11	21	38	70	81	93	89	100	9.92
	Girls	30	10	32	50	74	90	78	79	9.75
Giving rhymes	Boys	8	21	25	50	74	89	94	90	10.08
	Girls	13	13	36	45	74	90	83	77	9.92
Defining "fork" etc.	Boys	38	62	50	55	48	25	17	80	8.35
	Girls	61	65	61	68	39	20	11	81	8.09

actual correspondence. The greatest difference is that of 0.6 yr. in the test of giving the date. The differences, on the whole, are small, but must be taken into consideration when comparing the percentages that all boys and girls pass each test.

Two methods are available for studying the influence of sex differences on the individual tests. The first is that of comparing the results of boys and girls of each age on each test. This method is affected by the chance selection of superior or inferior subjects, and the results can have no meaning unless the relations of the groups of each age of the same sex are understood. For example, the fact that the 12 year boys are 36% lower than the 12 year girls in the test of naming the months has no significance as an isolated finding, for its significance is modified by the additional fact that this group of 12 year boys is 10% lower than the 9 year boys, 12% lower than the 10 year boys, and 9% lower than the 11 year boys on the same test.

The other method is that of comparing the per cent. that all subjects of each sex pass each test. This method avoids the factor of variations in the results due to a chance superiority of one age group over the other of the opposite sex, but, at the same time, it tends to obscure the magnitude of the differences that might occur. The most reliable differential measure between two groups is one that is well within the range of ability of the groups. The difference will be obscured if the measure is too easy or too difficult. A comparison of the results of all subjects would, in this way, tend to minimize² the magnitude of the real difference between the groups. Furthermore, there is a possibility that one sex might acquire an ability first, but eventually be surpassed by the other. The per cent. that all subjects passed would show no deviation, because the two tendencies would balance.

² For example, if there were 20 subjects of each age and of each sex from 6 to 12, and a certain test were passed by 75% of the 6 year girls, and by all of the 7, 8, 9, 10, 11 and 12 year girls, by 50% of the 6 year boys, 75% of the 7 year boys and all of the remaining groups, the total percentage passed for all girls would be 96%, and for all boys, 89%. The differential character of the test is indicated by the value 7%, while its actual differential character, just within the range of ability of the groups, is 25%.

Neither method, then, is entirely satisfactory, the first because it would tend to exaggerate chance differences, the second because it would tend to obscure real differences. The method used in this study is that of comparing the results of non-selected and selected subjects of each age and sex, studying first the general growth of each ability from age to age within each sex, and using the per cent. that all subjects pass each test to determine the correlation between the results of non-selected and selected subjects.

Table 13 shows the percentage of proportion³ that the boys and girls of each age pass each test, the percentage that all boys and girls pass each test, the actual percentage that the boys are superior to (+) or inferior to (—) the girls of each age, the difference between the average age of all boys and girls to whom each test was given, and the difference between the percentage that all boys and girls pass each test.

The differences between the performance of the boys and girls at each age have no meaning unless the general growth of the abilities in each sex is first understood. Studying first the results of the 187 non-selected boys shown in the first seven columns of Table 13, it may be seen that the growth of ability in each test is rather irregular. The test of naming the months, for example, shows a slight decrease from 9 to 12. The differences between the percentage performances of the subjects of each age and those of the preceding age were calculated. The 12 year group, compared to the 11 year group, is +11% on the test of giving the date, —9% on the test of naming the months etc. 61 differences were thus obtained, varying in magnitude from —15% to +36%, the median being +8% ($Q=9.75\%$). 13 of the deviations (21%) were minus values. The largest negative deviations occurred in the tests of naming colors (—15%, 7 to 8), naming money (—15%, 11 to 12), and constructing a sentence containing two ideas (—13%, 8 to 9). The remaining 10 minus deviations were less than 10%.

³ The proportion given is the number of times a test was given over the number of times a test was passed. No percentages were calculated for tests given less than 12 times, and no percentages are given for the definitions tests on account of the small number of times they are given to all subjects.

Percentage or Proportion that the Boys and Girls of Each Age Passed Each Test, and the Differences between the Percentages Passed.

	Chronological Ages												% passed all subjects	Differences in percentage passed at each age						Difference in average ages	Difference in all
	6	7	8	9	10	11	12	6	7	8	9	10		11	12						
Counting 13 pennies	Boys 84	93	94	8/7	7/7	3/3	1/1	90	-16	-4	-6					-23	-9				
Describing pictures	Girls 100	97	100	9/9	6/6	2/2	2/2	99	-1	+17	+10					-28	+5				
Copying a diamond	Boys 86	100	94	9/9	8/8	3/2	1/1	93	-21	+17	+22	+13				-44	+1				
Naming four colors	Girls 87	83	84	9/9	7/7	4/4	2/2	88	-21	+17	+22	+13				-27	-14				
Counting from 20 to 0	Boys 57	93	93	92	10/9	4/2	3/3	79	-30	+3	-19					-07	-8				
Counting stamps	Girls 78	76	71	79	8/7	4/4	3/3	78	-25	-7	+14	-9	+4			0	+6				
Repeating all digits	Boys 70	96	81	9/9	7/7	3/1	1/1	83	+14	+7	+7	-3	+21			-03	-6				
Naming days of week	Girls 100	93	100	9/8	7/7	3/3	2/2	97	-9	-7	-9	+9	-6	+1		-06	-16				
Giving day and date	Boys 4	25	55	50	79	6/3	8/7	42	-24	-16	-8	-12	-5			-60	-1				
Naming the months	Girls 29	32	41	59	75	7/6	5/5	50	-4	-7	-6	-7	-8	-11	0	+06	+4				
Copying designs from memory	Boys 25	29	62	75	79	6/3	8/7	52	-9	-7	-9	+9	-6	-7	+1	-10	0				
3 words in sent.	Girls 11	22	55	78	58	7/5	7/6	46	-24	-16	-8	-12	-5			-06	-16				
Two ideas	Boys 12	28	35	56	59	63	73	43	-4	-7	-6	-7	-8	-11	0	-13	-13				
3 words in sent.	Girls 21	35	44	47	65	70	72	49	-9	-7	-9	+9	-6	-7	+1	+10	+4				
One idea	Boys 35	59	75	88	95	6/5	8/8	67	-24	-16	-8	-12	-5			+06	+4				
Naming 60 words in 3 minutes	Girls 59	75	83	100	100	7/7	5/5	83	-4	-7	-6	-7	-8	-11	0	+10	0				
Giving rhymes with 3 words	Boys 13	18	48	84	92	89	100	64	+2	-23	+6	-18	-19	-36		+43	+3				
Defining by use	Girls 17	25	54	91	100	100	100	64	+11	+17	0	-1	+3	-18		+43	+18				
Defining superior to use	Boys 7	17	42	74	76	73	64	51	-4	-7	-6	-7	-8	-11	0	+17	+4				
	Girls 9/2	15	65	68	94	92	100	64	+11	+17	0	-1	+3	-18		+43	+18				
	Boys 10/2	11	38	33	60	72	57	44	-9	-7	-9	+9	-6	-7	+1	+17	+4				
	Girls 10/1	0	21	33	61	60	75	40	-4	-7	-6	-7	-8	-11	0	+17	+4				
	Boys 6/0	9/3	25	35	24	36	29	28	+2	-23	+6	-18	-19	-36		+17	+4				
	Girls 6/0	6/0	25	11	36	44	50	28	+11	+17	0	-1	+3	-18		+17	+4				
	Boys 3/1	9/3	80	67	74	73	78	70	-4	-7	-6	-7	-8	-11	0	+17	+4				
	Girls 6/0	8/4	63	63	65	84	88	67	+6	+1	+15	+26	+18	+18		+17	+4				
	Boys 3/1	9/3	27	22	45	58	67	44	+6	+1	+15	+26	+18	+18		+17	+4				
	Girls 6/0	8/0	21	21	30	37	41	26	-19	+7	-6	+19	+4	+4		+17	+4				
	Boys 4/0	6/3	9/4	36	60	77	69	58	-19	+7	-6	+19	+4	+4		+17	+4				
	Girls 7/0	3/0	9/6	55	53	83	50	54	-19	+7	-6	+19	+4	+4		+17	+4				
	Boys 3/0	6/4	6/5	10/4	57	60	65	58	-19	+7	-6	+19	+4	+4		+17	+4				
	Girls 3/1	4/3	10/8	10/7	59	61	80	68	-19	+7	-6	+19	+4	+4		+17	+4				
	Boys 14/13	18/16	12/12	11/10	15/15	7/6	3/3	94	-19	+7	-6	+19	+4	+4		+17	+4				
	Girls 14/13	20/15	17/16	15/14	9/9	4/4	2/2	90	-19	+7	-6	+19	+4	+4		+17	+4				
	Boys 14/13	18/6	12/7	11/6	15/3	7/2	3/1	33	-19	+7	-6	+19	+4	+4		+17	+4				
	Girls 14/3	20/4	17/5	15/5	9/3	4/1	2/0	26	-19	+7	-6	+19	+4	+4		+17	+4				

An index of the growth from year to year was obtained by calculating the average percentage increase from one age group to another. For example, the 7 year boys were 26% higher than the 6 year boys in the test of naming colors, 5% higher in naming the date etc. The average of the 10 possible comparisons between 6 and 7 year boys shows that the latter averaged 16.1% higher than the former. The average increases in percentage passed from year to year are as follows,—6 to 7=16.1%; 7 to 8=13.5%; 8 to 9=8.7%; 9 to 10=11.2%; 10 to 11=6.0%; and 11 to 12=0.2%. These figures show strikingly the irregularity of the growth from age to age. Comparing these average percentage increases in tests with the averages shown in Tables 9 and 11, there is no observable relation between this increase and the increase in average age from age to age, or the increase in average grade from age to age. The smallest increase in the tests (0.2%, 11 to 12) coincides with the smallest increase in average age from year to year (0.84 yr.), and the smallest increase in average grade from year to year. The other relations are varied.

The fact of the variability in the results of the non-selected boys stands out. The irregularity of the growth of the various abilities, and the fact that in 21% of the cases the boys of one age are actually lower than those of the previous age, point to the conclusion that certain allowances will have to be made for chance variations. It is not possible to account for the variations in growth by reference to the relative increase in average age or average grade from year to year.

The results of the 165 non-selected girls, shown in italics in the first seven columns of table 13, were studied in the same manner as the results of the boys. 60 differences between the percentage performance of the girls of each age and those of the preceding age were obtained. These differences ranged from —33% to +50%, the median being 7% ($Q=8\%$). 10 of the deviations (17%), were minus values. The largest deviations were shown in the tests of naming 60 words, (—33%, 11 to 12), counting stamps (—20%, 9 to 10), and drawing designs

(—14%, 8 to 9). The remaining 7 minus deviations were below 10%.

The average increases in the percentage passed from year to year are as follows,— 6 to 7=3.9%; 7 to 8=15%; 8 to 9=8.8%; 9 to 10=10.1%; 10 to 11=8.7%; 11 to 12=1.8%. Both boys and girls show the smallest average increase in the percentage passed in the step from 11 to 12, and the magnitudes of the increases agree fairly well except for the step from 6 to 7. The increase of the 7 year girls over the 6 year girls is 3.9%, the next to the smallest increase of one age group over any preceding group. The 7 year boys, however, show an average increase of 16.1%, over the 6 year boys, the largest increase of any group of boys over any preceding group. It will be difficult, then, to draw conclusions concerning sex differences from a comparison of the 6 year boys and girls, for the 6 year girls are either a superior selection or the 6 year boys are an inferior selection if the character of these groups be judged by the comparison with the 7 year subjects. The same comparison, on the other hand, might indicate that the 7 year girls were an inferior selection and the 7 year boys a superior selection from the general run. It is only possible to point out the irregularity, however, it is not possible to show the cause of the irregularity.

A comparison of the average increase in the percentage passed by girls from age to age with the increase in the average ages shown in Table 9 shows no demonstrable relation to exist. Comparing this growth in the ability on the tests with the growth in average grade, shown in Table 11, shows a very positive relation to exist between these factors. Where the increase in average grade is smallest (i.e. from 6 to 7 and from 11 to 12), the increase in the tests is smallest (3.9% and 1.8%), while the greatest increase in grade (from 9 to 10 and from 7 to 8) coincide with the greatest increase in the test abilities (10.1% and 15.0%). This relation was not indicated in the results of the boys. The explanation of this fact that a correlation between the increase in the tests with grade was found in the results of the girls but not of the boys is a matter of speculation. It has been shown that the boys have a higher variability in grade than

girls. This tendency of the boys to be distributed in a wider range of grades might nullify the grade correlation slightly, but probably not to any considerable extent. The fact that the causes of this variation are not determined serves to illustrate the dangers of comparing the results of two groups when the factors operating on the groups are not known.

The foregoing study of the growth of the various abilities from age to age in each sex, and the analysis of the causes influencing this growth, demonstrates the great variability of the results. This fact of variability must be considered before drawing conclusions concerning sex differences by the method of comparing the results of boys and girls of each age.

The percentage differences between the performance of non-selected boys and girls of each age are shown in Table 13. In actual magnitude, these differences vary from 0% to 36%, the median being 9% ($Q=5.5\%$). 75% of the differences are 17% or under, and only 16% are over 20%. In regard to sign, the differences vary from -36% to $+26\%$, the median being -3.5% ($Q=8.75\%$), showing a slight general superiority of the girls. If the number of possibilities of variation in comparing the results of small groups of non-selected subjects are taken into consideration, the presence of mental defectives, of subjects having language difficulties, of subjects in different grades influenced by different training, the possibility of a superior selection of subjects at one age group than at another, and the probability that similar chance samplings would not fall at the same age, the fact of correspondence indicated in Table 13 has more meaning than the fact of divergence.

The variability indicated in the study of the growth of abilities with age was so great that it makes interpretation of the results in terms of sex differences very difficult, and warranted conclusions impossible. It is legitimate to expect that the older subjects of either sex should make higher scores than the younger subjects of the same sex, but this was not found to be the universal rule. The boys' results showed minus deviations in 21% of the cases and the girls' results showed minus deviations in 17% of the cases. In one case the 12 year girls were 33% lower than

the 11 year girls. If this value (33%) be taken as the error due to chance variation, then only one value, that of —36%, (naming the months, age 12), may be taken as significant, and it has been seen that in this test the 12 year boys are 10% lower than the 9 year boys. The conclusion would follow, then, that there were no sex differences. This alternative, however, seems to place too much weight on one variation so that the truth probably lies in the assertion that the sex differences, that actually exist, are slight.

A study of the reactions of selected groups of boys and girls should throw light on the results from non-selected subjects, and make conclusions more certain. Subjects were selected by a process of elimination and selection. All of the subjects that were in the special class and minus grades were eliminated, along with all children of non-English speaking parents. From the following group of English speaking subjects in the regular grades all subjects were eliminated who had entered grade at an age very much above or below that of the general run of entrants.⁴ The remaining subjects ranged in age from 4.3 years to 14.4 years, but were found to group rather closely around certain ages. It was possible to find four groups of boys and girls of approximately the same chronological ages. The character of these subjects is indicated in Table 14.

The four groups of subjects, chronologically from 6.0 to 6.9, 7.6 to 8.9, 9.7 to 10.9 and 11.7 to 13.3 (which will be referred to as 6, 8, 10 and 12), were distributed in approximately the same grades, and had approximately the same average age and average grade. The results of these groups are shown in Table 15, which is arranged to show all the facts for selected subjects that were given for non-selected subjects in Tables 12 and 13. The first four columns show the percentage that each test was given to each group. The next four columns show the percentage or the proportion that the subjects in each group passed each

⁴ The ages on entering each grade of the subjects retained were as follows,—Kindergarten = 4, 5 and 6; Grade I = 5, 6 and 7; Grade II = 6, 7 and 8; Grade III = 8, 9 and 10; Grade IV = 9, 10 and 11; Grade V = 10, 11 and 12; Grade VI = 11, 12 and 13.

TABLE 14

Age in Grade Distribution, Average Grade and Average Age of 167 Selected Subjects. 86 Boys and 81 Girls.

Age in Grade Distribution												
Age Group	Sex	K	I	II	III	IV	V	VI	TOTAL	Average Grade (M.V.)	Average Age (M.V.)	
6.0 to 6.9	Boys	5	13						18	0.72 (0.40)	6.52 (0.22)	
	Girls	3	13	2					18	0.89 (0.39)	6.53 (0.22)	
7.6 to 8.9	Boys		7	13	3				23	1.83 (0.51)	8.09 (0.38)	
	Girls		2	13	5				20	2.15 (0.43)	8.32 (0.38)	
9.7 to 10.9	Boys				6	12	2		20	3.80 (0.48)	10.37 (0.36)	
	Girls				9	7	5		21	3.81 (0.69)	10.14 (0.32)	
11.7 to 13.3	Boys					2	8	15	25	5.52 (0.58)	12.35 (0.55)	
	Girls					3	8	11	22	5.36 (0.64)	12.41 (0.46)	

test. Column A shows the total number of times each test was given to all boys and girls, Column B, the weighted average age (the average ages given in Table 14 being used), and Column C the percentage that all subjects passed each test. The next four columns show the percentage that the boys are above (+) or below (—) the girls. Column D (derived from Column B), gives the difference between the average ages of all subjects to whom each test was given. Column E (derived from Column C), gives the differences between the percentages passed by all boys and girls on each test.

The growth of the various abilities with age in the selected groups of subjects is more uniform than that shown by the non-selected subjects. Only three cases appear in which the younger subjects make higher scores than those of older subjects, these exceptions occurring in the tests of describing pictures (—3%, girls 6 to 8), naming colors (—7%, girls 6 to 8), and naming months (—9%, boys, 10 to 12). In the comparison of the sexes 41 differences are obtained varying in magnitude from —28% to +26%, the median being 0% ($Q=9.5\%$). In actual magnitude the differences vary from 0 to 28, the median being 10% ($Q=4.75\%$), the median being 1% higher than that of non-selected data, and the variability 0.75% less. 75% of the differences were less than 14%.

TABLE 15
Results of 167 Selected Subjects. (86 Boys and 81 Girls).

	Percentage test was given.				Percentage or proportion seen.				Columns			Percentage that boys are higher or lower than girls				Columns	
	6	8	10	12	6	8	10	12	A	B	C	6	8	10	12	D	E
Counting 13 pennies.	100	87	15	0	94	95	3/3		41	7.57	95	-6	-5			-25	-5
Describing pictures.	100	70	24	5	100	100	5/5	1/1	38	7.82	100	+11	+14			-44	+10
Copying a diamond.	100	83	20	0	100	100	4/4	1/1	41	7.38	100	-27	+8			-16	-8
Naming four colors.	100	83	40	0	89	86	5/5		38	7.82	90	-11	+2			-27	-4
Counting from 20 to 0.	100	60	43	5	83	92	9/9	1/1	45	7.87	82	-27	-28	+2		+10	-16
Counting stamps.	100	83	15	0	89	95	3/3	1/1	40	8.03	90	+1	+8	+26		-08	+11
Repeating all digits.	100	70	24	5	100	93	5/5		38	7.82	97	-11	0	-11	+4	+01	-4
Naming days of week.	100	96	75	12	6	36	93	3/3	57	8.45	46	-23	+9	0		+14	-2
Giving day and date.	83	70	52	9	33	64	9/1	2/2	42	8.35	62	-12	-12	0		+03	-5
Naming the months.	100	100	75	12	14	48	93	3/3	55	8.50	55	-23	+9	0		-20	-12
Naming pieces of money.	89	75	57	23	13	40	67	5/5	48	8.58	44	-11	0			+03	-2
Copying designs from memory.	100	100	100	100	16	39	56	7/5	86	9.53	48	-11	0			+08	-1
3 words in sent. two ideas.	100	95	100	100	27	39	67	7/1	80	9.52	52	-23	+9	0		+40	-2
3 words in sent. one idea.	100	100	100	100	8	27	100	100	67	9.37	60	-12	-12	0		+40	+18
Naming 60 words in 3 minutes.	67	96	90	68	20	39	100	100	66	9.34	65	-12	-12	0		+40	+18
Giving rhymes with 3 words.	83	90	86	68	10/1	29	94	85	61	9.38	56	-7	-7	0		+49	+1
Defining by use.	56	91	85	52	7/2	41	88	100	52	9.58	67	-12	-12	0		+29	+6
Defining superior to use.	39	85	76	55	9/1	31	52	82	60	9.85	50	-12	-12	0		+29	+6
	50	65	100	77	2/0	14/3	30	53	53	10.26	34	-7	-7	0		+29	+6
	11	61	100	68	5/0	9/3	24	63	51	10.18	35	-7	-7	0		+29	+6
	28	45	100	73	1/1	13/6	75	88	59	10.64	75	-7	-7	0		+29	+6
	6	57	100	100	6/0	10/9	71	100	57	10.24	77	-7	-7	0		+29	+6
	33	50	100	91	1/1	13/2	50	68	57	10.64	51	-7	-7	0		+29	+6
	33	50	100	91	6/0	10/3	33	45	57	10.24	33	-7	-7	0		+29	+6
	11	30	80	96	2/0	7/3	69	88	47	10.33	53	-7	-7	0		+29	+6
	30	25	76	86	7/0	5/3	56	68	47	10.33	53	-7	-7	0		+29	+6
	6	22	65	96	1/0	5/3	69	71	43	11.12	67	-7	-7	0		+29	+6
	17	30	76	86	3/1	6/4	63	74	44	10.63	66	-7	-7	0		+29	+6
	44	65	55	16	8/8	15/15	11/11	4/4	38	8.87	100	-7	-7	0		+29	+6
	61	55	43	14	11/10	11/10	9/9	3/3	34	8.58	94	-7	-7	0		+29	+6
	44	65	55	16	8/1	15/9	11/4	4/4	38	8.87	47	-7	-7	0		+29	+6
	61	55	43	14	11/2	11/2	9/5	3/0	34	8.58	26	-7	-7	0		+29	+6

The change of the median of the series of differences from -3.5% (non-selected) to 0% (selected) shows that the elimination of over age and special grade pupils has helped the boys more than the girls, and has altered the general relations between the sexes. This fact is also indicated by the average difference in the percentages that all subjects pass each test, the average for non-selected subjects being -1.4% and for selected subjects $+1.6\%$. The non-selected boys from 6 to 12 were given, in all, 2436 tests, these tests being passed 60.8% of the time. The non-selected girls were given 2195 tests, passing 61.6%, the advantage being 0.8% in their favor. The selected boys were given 1125 tests, passing 64.3%, an advantage of 0.1% over the girls who passed 64.2% of 1034 tests. The foregoing changes indicate clearly that the selection of subjects has changed the general relations between the sexes, helping the boys more than the girls.

The relations between the results of selected and non-selected subjects may be studied by a comparison of the differences between the percentages passed by all subjects. If the differences between the scores of the boys and girls are due to but one factor, that of sex differences, then the correlation between the two methods of study should be very nearly absolute. The correlation (Pearson product-moments formula) between the differences in the percentage passed by all boys and girls according to the two methods is 0.726 ($p=0.075$). This correlation between the two methods is high, but it would probably be high inasmuch as the 167 selected subjects are included in the 352 non-selected subjects. The results of the two methods show certain large discrepancies. The changes of the greatest magnitude are those shown by the 60 words test ($+4\%$ by the first method to $+18\%$ by the second), the tests of defining in terms superior to use ($+7\%$ to $+21\%$), of naming the days of the week, (-16% to -2%), giving rhymes, (-10% to $+1\%$), naming colors, (-14% to -4%), copying the diamond, ($+1\%$ to -8%), and counting from 20 to 0 (-8% to -16%). The comparison of the median differences shows that the selected method tends to improve the results of the boys more

than those of the girls. All of the changes in the results of the two methods are not in favor of the boys, however, the total scores on the diamond and 20 to 0 tests showing changes in favor of the girls. If the cause of the variations shown by the first method is the presence of a few children of non-English speaking parents, to special class and minus grade children, then the elimination of this source of error should change the results in only one direction.

The analysis of the results of selected subjects, therefore, does not lessen the difficulty of the interpretation of the results in the light of sex differences. The rate of growth of the various abilities with age is irregular. The analysis of the irregularities points to the fact that the boys or girls of any age may be a chance selection of superior or inferior subjects at that age. The method of comparing selected subjects would tend to eliminate the inferior selection of subjects, but would not eliminate the possibility of a superior selection.

The comparison of the results of the sexes shows differences at certain ages and on certain tests that are as high as 20%. The problem involved is that of deciding whether these large differences are due to chance or to differences in the reactions of the sexes. Certain tests show large deviations first in favor of one sex and then in favor of the other. If a difference of a percentage of any magnitude on any test is to be attributed to a sex difference, then the same line of reasoning will show that in certain tests the abilities change from one sex to the other. The analysis of the tests that show this crossing of ability should throw light on the other tests.

Three tests show substantial differences in favor of both sexes according to both methods. In the test of copying the diamond, the non-selected girls lead at the start, age 6, and the boys are ahead at 7, 8 and 9, the same relations being shown by selected subjects of 6 and 8. In the test of copying the designs from memory, the non-selected girls are 24% below the boys at age 9 and 21% above the boys at age 12, the same relations being shown by the selected subjects of 10 and 12. In the test of naming 60 words in three minutes, the non-selected girls are

19% above the boys at 9, and 19% below at 12. The selected boys of 10 and 12 are in advance of the girls in this test.

These three tests are crucial in the consideration of the problem of whether differences shown between the boys and girls are due to actual sex differences or due to accidental causes. Each of these tests may be studied by a method more accurate than that of comparing the percentage passed at each age. The reproductions of the diamond were arbitrarily sorted in six groups according to their merits by a method described in the discussion of the personal equation. The first group contained the best reproductions, the sixth, the poorest. The reproductions of the designs were graded from 0 to 20 by an arbitrary point system described under the discussion of the personal equation. A measure of the ability in the 60 word test is the actual number of words given in three minutes, a measure recorded by the experimenters in each case. Table 16 shows the average score made by the non-selected and selected boys and girls of each age in these three tests.

TABLE 16

Average Score (Mean Variation) of Subjects of Each Age on Three Tests.

	Copying the Diamond Average Group of the Reproductions.		Drawing the Designs Average number of points scored.		Naming 60 words Average number of words given in three minutes.	
	Boys	Girls	Boys	Girls	Boys	Girls
unselected subjects	6	4.27(1.28)	3.57(1.24)			
	7	2.85(1.04)	3.17(1.37)			
	8	2.20(1.15)	3.24(1.57)	8.06(6.19)	9.00(5.25)	
	9	2.33(0.89)	3.00(1.29)	10.29(5.30)	5.32(4.61)	52.93(11.20)
	10			9.17(5.33)	9.18(6.73)	68.12(13.12)
	11			8.64(6.73)	10.94(7.06)	73.65(13.35)
	12			8.64(6.02)	11.08(6.08)	68.75(12.28)
selected subjects	6	4.27(1.20)	3.33(1.26)			
	8	2.32(1.00)	3.00(1.17)			
	10			9.55(5.60)	7.29(6.42)	67.31(12.74)
	12			12.53(5.38)	13.56(5.55)	62.13(11.39)
						75.33(10.92)
						66.84(13.87)

The relations indicated by the percentage passed are also indicated by the more reliable method of comparing the average scores. In the test of copying the diamond, the 6 year non-selected girls average 0.70 group better than the boys, while the

selected girls are 0.94 ahead. The comparison of the 7, 8 and 9 year subjects shows the boys ahead in all cases, the 8 year non-selected boys averaging over one group higher. The non-selected boys show an improvement of two groups from 6 to 9, while the girls show an improvement of only half a group. One sex shows a decided growth of ability, the other practically none. If the differences indicated are to be taken as real, it will be necessary to assume that the girls pick up the ability to draw a diamond easier than the boys, but that this ability once obtained remains constant—that the effect of maturity operates on one sex but not on another. The number of cases on which this assumption is based (174 subjects from 6 to 9) is so small, and the chances of variation in the selection of subjects of different intellectual status in each age group is so large, that the assumption is not substantiated.

The relations indicated in the test of copying the designs are more variable than those of the diamond test. The 9 year non-selected boys show an improvement over the 8 year boys, but from 9 to 12 there is a gradual decrease in the ability, so that the 11 and 12 year boys are only slightly ahead of the 8 year boys. The relations shown by the non-selected girls are exactly the reverse of those of the boys. The 9 year girls are very much lower than the 8 year girls, and a gradual increase appears from 9 to 12 instead of a decrease. The comparison of these opposite relations gives a maximum difference in favor of the boys at 9 and the girls at 12. If the relations indicated in this test are to be considered definite, the assumption is involved that the influence of increasing age on one sex is exactly opposite to that on the other sex, an assumption that is not substantiated in view of the small number of cases (183 subjects from 8 to 12) and the possibility of selecting subjects of chance superiority in the small groups at each age.

The relations indicated in the test of naming 60 words are more constant than those shown in the diamond or design tests. Both sexes show a growth of ability from 9 to 11 and a decrease from 11 to 12. The growth is irregular, however, the girls showing less growth from 9 to 10, and a greater drop from 11 to

12, so that a comparison of the sexes shows a deviation in favor of the girls at 9 and of the boys at 12. The assumption of any large sex differences in this test involves the assumption that 12 year girls have less ability in this test than 9 year girls, and that the influence of maturity operates differently on the two sexes, an assumption that is not substantiated in view of the many variable factors.

The conclusion that a definite crossing of ability between the sexes occurs in the tests of copying the diamond, copying designs and naming 60 words, is not substantiated. It is not justifiable to attribute a difference of 20% between the sexes to a real sex difference on one test and not on another. If the differences shown between the results of the sexes in the tests of constructing a sentence containing one idea, of naming the months, naming the days of the week, counting stamps and naming colors are to be attributed to sex differences, then the variations in ability shown in the diamond, design and 60 word test must be assumed to be definite. These assumptions were not found to be substantiated, however, so that it is not possible to draw any conclusions concerning sex differences from a study of the percentage that selected or unselected subjects of each age pass each test.

The variable influences due to the selection of subjects of different status at each age are eliminated or counterbalanced to some extent by combining the subjects of all ages. The differences between the percentages that all boys and girls pass each test are to some extent influenced by the ages of the subjects to whom each test was given. The correlation (Pearson product-moments formula) of the differences between the percentages that all non-selected boys and girls passed each test with the difference between the average ages of all the non-selected boys and girls to whom each test was given is 0.394 ($p=0.134$). The correlation between the same arrays from selected subjects (i.e. between Columns D and E of Table 15) is 0.388 ($p=0.135$). These correlations between the tests and age are high enough to indicate that the factor of age is present to some extent. The close correspondence in the correlations from the two methods

indicates that the age factor is present to the same extent in both methods. The tests vary in the degree with which they correlate with age, so that it is not possible to estimate the amount of the influence of this factor. Furthermore, it has been seen that the results from the two methods are not in strict accordance, that the elimination of inferior subjects caused changes in the results in both directions. For these reasons, it is not possible to draw any conclusions concerning sex differences from a comparison of the percentages passed by all subjects.

Certain negative conclusions are, however, possible. The number of subjects at each age in both methods is comparatively small. The chances of variations due to factors other than sex differences has been shown to be very large. The fact of correspondence between the results of the two sexes is therefore of more importance than the fact of divergence. 75% of the differences between the non-selected boys and girls are 17% or under, while the same proportion of the differences between selected boys and girls falls under 14%. If it is assumed that the subjects of any age should not test lower than those of any preceding age, and allowance is made for differences between the sexes that are exaggerated on account of the chance falling off of ability with older subjects, only 9% of the differences between the non-selected boys and girls are over 20% (derived from Table 13).

The evidence from the foregoing methods of study points to the conclusion that the sex differences, if present, are under 20% or 25% as a maximum, and that deviations of this magnitude are marked exceptions to the general run of differences. The conclusion that the differences that might possibly be attributed to the sex factor are slight, has no meaning unless the word "slight" is defined independently of the writer's personal opinion. The differences shown between the results of the sexes are smaller than those that were attributed to the factor of the personal equation in the study of the results of the four experimenters. It was concluded that certain tests were influenced by grade training. These tests showed from 40% to 60% improvement from one grade to another, so that the greatest influence that may be attributed to the sex factor is only approximately

one half that due to grade training. The following study of the diagnostic value of the tests will show that the deviations that might be attributed to the sex factor are insignificant when compared to the differences between the reactions of normal and retarded children to the individual tests.

Most of the investigators who have studied the factor of sex differences in the Binet tests, have studied them from the standpoint of the "mental ages" or total scores made by the subjects of both sexes. A few investigators have studied sex differences in the light of the individual tests. Descoeudres (20) reports the results of the application of the Binet tests to 24 subjects, one good and one poor pupil of each sex from each of six school grades, drawing conclusions from this investigation concerning the diagnostic value of the individual tests and the sex differences involved. Obviously the number of subjects is too small to allow any conclusions to be drawn. Chotzen (18) compared the percentage that all feeble-minded boys and girls passed each of 15 tests, finding differences varying in magnitude from 1% to 20%. The largest deviations were those of 20% in favor of the boys in the test of copying the diamond, 13% in favor of the girls in the test of executing three commissions, 12% in favor of the boys in naming the pieces of money, 11% in favor of the girls in the test of repeating a sentence of 16 syllables, and 10% in favor of the girls in detecting omissions in pictures. All other differences were less than 10%.

Bloch and Preiss (9) examined 155 normal *Volkschule* children (79 boys and 76 girls) varying in age from 7 to 13. Bober-tag's translation was used. These investigators found very striking differences in the reaction of the sexes to the individual tests, the differences running as high as 52%, most of them in favor of the boys. The differences between the performances of the boys and girls of each age were calculated, without reference to the many sources of variation. The factor of the personal equation is not treated, and this factor alone might cause these variations. If a more careful analysis of the results had been made, it is very probable that the conclusions would have been modified to some extent. The fact that the 11 year

boys are 37% higher than the 11 year girls on the test of criticising absurdities is most certainly modified by the fact that the 11 year subjects are 30% lower than the 10 year subjects in the test of repeating 7 digits. The small number of subjects (in five cases less than 10), would tend to emphasize chance variations. The fact that the number of subjects is too small to warrant definite conclusions is pointed out by the authors. Stern (62) in commenting upon these results, points out the significance of the fact that the inferiority of the girls extends to so many different kinds of tests. The results of Bloch and Preiss are in almost complete contradiction to the results of the present investigation. They find large differences, and find practically all of these differences in favor of the boys. This investigation shows a general run of differences very much smaller, and a slight general superiority of the non-selected girls. The mere fact of contradiction in the results of the two investigations would indicate that the differences were not produced by the common factor of sex. Rogers and McIntyre (54) give no figures, but report that they have studied their results in the light of sex differences, and have found no correlation between their results and those of Bloch and Preiss.

The results of the investigators who have compared the "mental ages" or total scores of children of different sexes are somewhat at variance. Goddard (30) reports that there are more backward boys than girls. Stern notes that Goddard's results do not bear out his statement, for the percentage of boys and girls testing two or more years retarded is the same (18.5%). The accuracy of Goddard's statement depends on the criterion⁵ used for measuring backwardness. Although Goddard's state-

⁵ If the criterion is four or more years retarded, there are more backward boys than girls (boys = 3.7%, girls = 3.1%). If the criterion is three or more years backward, there are more girls than boys (boys = 8%, girls = 9.1%). If the criterion is two or more years backward, the proportions are the same, as Stern notes. If the criterion is one year or more retarded, there are more backward boys than girls (boys = 41.4%, girls = 35.6%). There are more girls than boys testing at and above age according to Goddard's results. 34.7% of the boys and 36.6% of the girls test at age, while 23.8% of the boys and 27.7% of the girls test one year or more above age.

ment concerning the backwardness of the boys may be interpreted differently, his figures leave no doubt concerning the fact that there are more girls than boys at and above age, and therefore indicate a general superiority of the girls.

Bobertag (10) computed the average "mental age" of 90 boys and 90 girls regularly distributed from 7 to 12. The subjects were selected according to school grades, so that the average grade of each group differed by exactly one grade. His results show the boys ahead 0.06 yr. at 7, 0.14 yr. at 8 and 9, 0.20 yr. at 10, 0.19 yr. at 11 and 0.14 yr. at 12. These findings cannot be considered entirely out of harmony with those of Goddard, for, as this investigation shows, there may be a change in the relation of non-selected boys and girls and selected boys and girls.

Yerkes and his co-workers (82), scoring some of the Binet tests according to the point system, show that the girls of English speaking parents are superior to the boys of the same parentage between 5 and 7, that they fall below with minor variations till 11, where they again surpass the boys at 12 and 13, falling below at 14 and 15. The differences between the sexes are smaller and of less practical importance than the differences due to the language factor, but the authors suspect "that at certain ages serious injustice will be done to individuals by evaluating their scores in the light of norms which do not take account of sex differences." (page 73). In contradiction to these results are those of Terman and his co-workers (67), who, scoring the Stanford revision of the Binet scale according to "intelligence quotients," find differences of but 2% to 4% in these quotients in favor of the girls, and who conclude from the basis of their studies of sex differences that the conclusions of Yerkes are unjustified. These two investigations used tests different in character and differently weighted, so that the results would not necessarily have to correspond.

The one common feature of most of the researches on sex differences in the Binet-Simon tests is that the differences are small. Burt and Moore (17) summarize the work of various investigators in the general field of sex differences, and report an investigation of their own on 67 boys and 63 girls, 12½ to 13½

years of age. They discuss their results and those of the other authors in the order of the complexity of the mental processes involved. They find a high correlation between the size of the sex difference and the simplicity of the capacities compared—the higher the process, and the more complex the capacity, the smaller the sex difference.

The general trend of the investigations on sex differences indicates that no very large differences are to be expected in the application of intelligence tests, and that the differences to be expected will vary according to the nature of the tests. The results of this investigation are in agreement with the general trend of the investigations in showing only slight differences that might be attributed to the sex factor. The results do not show on what tests, if any, these differences occur. Conclusions concerning the amount of influence of this factor must be drawn from more exhaustive investigations on the individual tests. The research of Bateman (3), for instance, is conclusive in the test of naming colors. Bateman shows that there is a difference of 14% in favor of the girls in this test, showing furthermore that the factor of school training causes an improvement of but 18%. The results would indicate that the test should be placed in the fifth or sixth year, but the sex difference of 14% would probably not warrant the placing of the test in a different age group for boys and girls.

The investigations of Bolton (11) and Wooley (79) would show that small differences in favor of the girls are to be expected in the tests of repeating digits, and possibly in all memory tests. The investigations of Gilbert (27), Thompson (68), Burt and Moore, and Peterson and Doll (51) would indicate that a slight difference in favor of the boys should appear in the test of arranging five weights. Ruger's (55) finding of striking differences in favor of men in a series of puzzle tests, and Wooley and Fisher's finding of large differences in favor of the boys in the Healy puzzle-box test would show that rather large differences might appear in the general class of "puzzle" tests.

Even though the sex differences in intelligence tests may be shown to be small, scientific procedure should demand that the

investigator who standardizes any test or system of tests should treat his results in such a way as to demonstrate that the factor is present or not present. The burden of proof should still be on the person who maintains that sex differences are not involved. The knowledge of sex differences is especially important in diagnosing border-line cases of mental defect, where the diagnosis must often be made on the qualitatively different character of the responses to individual tests.

VI. SUMMARY.

One of the fundamental assumptions in the construction of the Binet-Simon scale is the correlation of the individual tests with age. The correlation of the tests with age is affected by the error due to incomplete data, by the influence of the personal equation of the experimenter, and by the training the subject has received in school.

The influence of the personal equation of the experimenter was found to be more marked in some tests than in others, the influence being most marked in the tests of copying the diamond, indicating omissions in pictures, defining in terms superior to use, drawing designs from memory, detecting absurdities in statements and reconstructing dissected sentences.

The variations between the experimenters could be traced to three sources,—

- 1) to the use of apparatus, variations in which were due to,
 - a) the construction of the test material, and
 - b) the use of alternative questions;
- 2) to the technique of the experimenters in giving the tests; and
- 3) to observation errors made by the experimenters in marking a response passed or failed.

It is possible to eliminate all three sources of error.

The effect of school training was more marked on some tests than on others, the effect being most marked in the tests of counting stamps, counting backward from 20 to 0, enumerating the days of the week and the months, giving the day and the date, naming the pieces of money, making change, and reconstructing dissected sentences. Tests that involve school training should be standardized on a different basis than those relatively independent of this factor.

Although the comparison of "mental ages" and pedagogical ages gives no information concerning the general correlation be-

tween the Binet tests and the school grades, the study of the individual tests establishes the fact of a general correlation.

The correlation of the individual tests with grade is higher than the correlation of the tests with age, this fact being indirect evidence of the value of the tests as measures of intelligence.

Sex differences were found to be slight as compared with the influence due to the personal equation or grade training.

Since variations occur in the results due to the influence of the personal equation and grade training, certain allowances must be made for these factors in making diagnoses on the basis of the tests. The scale is therefore a qualitative rather than a quantitative instrument.

The investigator who wishes to use his results for standardizing age norms should use only those data based on the complete method of experimenting, and should treat his results in such a way as to demonstrate the presence or absence of the variable factors of the personal equation, grade training and sex differences.

THE DIAGNOSTIC VALUE OF SOME MENTAL TESTS

TABLE OF CONTENTS

I.	Introduction	95
II.	Character of Subjects	106
III.	Tests and Procedure	116
IV.	Variable Factors	126
V.	Diagnostic Value of Binet Tests	140
VI.	Diagnostic Value of Supplementary Tests	166
VII.	Correlations of Abilities with Age	205
VIII.	Results of Other Investigators	222
IX.	Conclusions and Suggestions	232
	Bibliography	249

I. INTRODUCTION

The Binet scale has been used to classify normal children, and as a means of studying mental differences due to race, sex and environment, but its most important function is the detection and classification of conditions of feeble-mindedness. The consensus of expert opinion would seem to show that the scale is unsatisfactory in this respect. Among the questions brought up at the Buffalo conference (15) was the following,—“Does the scale provide a reliable means of diagnosing feeble-mindedness?” The answer given was,—“It does not always furnish a sharp, nor a positive diagnosis of feeble-mindedness: in particular—*a*. A mental age of 10 or above is not necessarily indicative of feeble-mindedness, regardless of how old the examinee may be; and *b*. A young child may test almost at age and yet be feeble-minded as determined by other criteria.”

W. E. Fernald (26) discussing the question of detecting the higher grades of mental defect, writes “The Binet tests, in the hands of competent examiners, usually corroborate the results of clinical examination in the recognition of all degrees of mental defect in children under ten, and of pronounced defect in older persons. These tests are not so effective in detecting slight mental defect in world-wise adolescents and adults. In other words, the Binet tests corroborate where we do not need corroboration, and are not decisive where the differential diagnosis of the high grade defective from the normal is in question” (page (747)). And again, “The Binet test does not register as defective certain persons who present plain evidence of mental defect in their personal history, school history and performance, social and economic relations, etc., while on the other hand, certain individuals who fail to come up to the requirements of the Binet test do not present the usual personal social and economic reactions of mental defect” (page 748).

These opinions are corroborated by the lack of agreement be-

tween investigators who have made studies of groups of adolescents. In the studies of delinquency for example there is a wide disagreement between investigators. M. Otis (50) examined 172 girls, ages from 10 to 20, in a state home for delinquent girls to which commitments were made by the courts, and reports 25% "presumably normal," 30% morons, and 45% defective, i.e., 75% feeble-minded. Morrow and Bridgman (47) report the results of an examination of 60 girls of a similar run of ages in a similar institution in a neighboring state, finding 10% normal, 66% feeble-minded and 24% "doubtful." Bridgman (13) reporting the results of the examination of 118 girls at the same institution, the general run of admissions, gives 5% normal, 6% backward and 87% feeble-minded. Of the 104 girls committed as sexual delinquents, Bridgman finds 3% normal and 97% feeble-minded. These investigators used the same instrument, Goddard's 1911 scale. Healy (33) working in the same general field with a much larger group of delinquents and with more reliable methods, finds a much smaller percentage of feeble-mindedness. In fact, his group of feeble-minded is but 11.3% of the total number of cases (823), and the entire group of defective types (including cases of feeble-mindedness, poor native ability, mental subnormality, dullness from physical causes including epilepsy, and specialized defects including defects of self-control) is but 25.3% of the whole group of delinquents studied.

That the disagreement between the various investigators in the proportion of feeble-minded individuals reported among delinquents is due to the tests is shown by Kohs' (42) study of 335 cases at the Chicago House of Correction. Kohs used Goddard's 1911 revision of the Binet scale. He also used other criteria for deciding whether his subjects should be classified as normal or feeble-minded. He found that the feeble-minded individuals tested from $6\frac{1}{5}$ to $11\frac{2}{5}$, the normal individuals from $10\frac{4}{5}$ to $12\frac{2}{5}$. In other words the results of the Binet scale, instead of showing a positive differentiation between normal and feeble-minded, showed a marked over-lapping of per-

formance. That this error in the scale in failing to make a complete differentiation is a serious one is shown by the fact that 30% of Kohs' cases fell within the range of over-lapping (10 4/5 to 11 2/5).

Diagnoses of feeble-mindedness will probably never be made on the basis of mental tests alone, but the reliability of these diagnoses will most certainly be increased if the tests are improved. The lower grades of feeble-mindedness cause little difficulty in diagnosis. Idiots are self-diagnostic, and imbeciles rarely reach adolescence without detection. The moron group is the most difficult to diagnose. Inasmuch as this group is also the most difficult to treat socially, it would seem worth while to perfect the instruments of diagnosis.

Competent authorities maintain that the differences between the normal individual and the idiot are not differences of quality or species, but differences only of quantity or amount of intelligence. On this theory intelligence will be found in varying amounts from idiocy to genius. As intelligence is a means of adjusting the individual to his environment, individuals will be found varying in the degree of adjustment from the idiot who can not feed himself to the individual competent to control his environment in a number of ways. Some authorities, for instance Witmer (74) would hold that the diagnosis of feeble-mindedness is a social diagnosis. According to Witmer, the diagnosis is not made concerning the subject's mentality, but is concerned merely with the advisability of freedom or segregation. This view would seem to confuse the disease and its treatment, for the defective social adjustment of the feeble-minded is always referred to defective intelligence, just as the defective social adjustment of the blind person is referred to his lack of eye-sight. The most profitable method of increasing the accuracy of the diagnosis of feeble-mindedness would therefore seem to be that of increasing the reliability of tests of intelligence.

If the Binet scale fails to detect the higher grades of mental defect, it is legitimate to ask why it fails. Binet (7) suggested the answer in discussing the means of differentiating the moron

from the normal. He considered that six tests(arranging five weights, comprehending difficult questions, using three given words in a sentence, defining abstract terms, interpreting pictures, and giving rhymes) were important in distinguishing the moron from the normal individual of the Paris population. Binet therefore considered certain tests more diagnostic of intelligence than others. However, he offered no proof that these six tests were better than any others and his assertion can be no better than an expression of opinion, with the possibility that his opinion was wrong. In the actual construction of the scale, the tests were weighted equally, the individual receiving the same amount of credit for passing the test of naming the months as he would for passing the comprehension questions, the test which according to Binet dissipated all his doubts concerning a final diagnosis. (See page 3.) Each test counts for one fifth of a year, and it makes no difference in the quantitative score whether a child reaches a certain "mental age" by passing the most diagnostic or the least diagnostic tests. The final diagnoses that Binet made in his own cases must have been qualitative rather than quantitative for he threw more weight on some tests than on others in forming his opinion.

Following Binet's cue on the matter, it would seem that the method of increasing the accuracy of the diagnosis of the higher grades of mental defect would be that of determining what sorts of tests were the most highly diagnostic of this sort of defect. In the study of the Binet scale itself, the problem of determining what tests are diagnostic of intelligence immediately arises. Binet demonstrated that the tests were correlated with age, but he did not go beyond this point. He never demonstrated that the tests were correlated with intelligence. In his opinion of course, all the tests were diagnostic of intelligence or he would not have included them in his "Measuring Scale of Intelligence," yet it is plain from his writings that he considered some tests more valuable than others in this respect. It remains for other investigators to check up Binet's work, and to establish the accuracy of the individual tests that he included in his scale. Studies of the

diagnostic value of the Binet tests have been made by Descoedres and Chotzen.

Descoedres (20) published in 1911 the results of the application of the scale to one good and one poor subject of each sex from each of six school grades, using the material from these 24 subjects as a basis of studying the relative diagnostic value of the tests, and of studying the sex differences involved. Grouping the subjects according to the teachers' judgments of good or poor, the tests that showed the most marked differences between the two intellectual levels were those of arranging five weights, interpreting pictures, detecting absurdities, defining in terms superior to use, counting backwards from 20 to 0 and indicating the omissions in pictures. The group with inferior intellectual endowment made lower scores on all tests except those of understanding easy and difficult problem questions.

Descoedres (21) also published in 1911 the results of an investigation on 14 backward and defective boys and girls from 6 to 14 years of age. 15 tests were used; 6 of them (describing pictures, defining, comparing remembered objects, naming as many words as possible in three minutes, comprehending questions, and recognizing coins) being taken from the Binet scale, the other tests being tests of motor dexterity, tactile ability, auditory and visual imagination, puzzle solving, cancelling a's, calculation, and auditory and visual memory. The obvious differences between the subjects made the independent (rank) estimation of their intelligence possible, so that the correlation of each of the tests with intelligence could be determined. In the list of 15 tests arranged according to the magnitude of their correlation with intelligence, comparing remembered objects showed the highest correlation, describing pictures and comprehending questions stood third and fourth, defining sixth, naming four pieces of money tenth, and naming words fifteenth. Analyzing the factors involved in each, Descoedres concluded that the tests of reasoning show the highest correlation with intelligence, imagination next, while tests of memory, particularly auditory memory, show the lowest correlation. Although these

two investigations of Descoeudres are suggestive, the small number of subjects precludes the possibility of forming any definite conclusions concerning the diagnostic value of the tests.

In 1912, Chotzen (18) published the results of the application of the Binet scale to 280 backward and defective children. The subjects, 157 boys and 123 girls varying in age from 7 to 14, were either enrolled in the special classes for backward and defective children in Breslau, or were candidates for admission to these classes. The largest proportion of the subjects were distributed in the ages 8, 9 and 10. Bobertag's standardization of Binet's 1908 scale were used, so that Chotzen could compare his results from backward children with those of Bobertag from normal subjects. The tests used were largely those in the V, VI, VII, VIII, and IX year groups. The tests for "ten years" were given to some extent, but those in the higher groups were very rarely used.

Chotzen subjected the results of his investigation to a very careful analysis, part of which deals directly with the relative value of the tests in the diagnosis of feeble-mindedness. He used three methods of analysing the data,—(1) the comparison of his results from feeble-minded children with those of Bobertag from normal children; (2) the comparison of the results of feeble-minded children of the same mental ages but different chronological ages; and (3) the comparison of the results of groups of children of approximately the same ages but with different final medical diagnoses.

In studying the results according to the first method, Chotzen found that the feeble-minded children were in general from 2 to 4 years backward as compared to normal children, the deficiency being more marked in some tests than in others. The test of describing pictures, for example, placed by Bobertag in year VII, was passed by 73% of Chotzen's feeble-minded children of 8 years, while the test of counting backwards from 20 to 0 in Age VIII was passed by only 8% of the 8 year, 16% of the 9 year, 29% of the 10 year, 55% of the 11 year and 72% of the 12 year feeble-minded children. The first test would be a "seven"

or an "eight year" test for feeble-minded children, while the second would be a "twelve year" test.

Backwardness was more marked, then, in the latter test. On this basis, the tests in which backwardness was most marked were those of repeating a sentence of 16 syllables, making change, counting backwards from 20 to 0, defining in terms superior to use, comparing remembered objects, recalling a story read, naming the months, repeating five digits and arranging five weights. Backwardness was least apparent in distinguishing between morning and afternoon, defining in terms of use, describing pictures, counting 13 pennies, giving age, choosing the prettier of given faces and counting the fingers.

The second method gave Chotzen an opportunity to study the effect of school training and physical maturity on the tests. According to this method, the results of children of the same "mental age" but of different chronological ages were compared. Taking all subjects of the "mental age" of eight, those of the chronological ages 8 and 9 passed the tests of repeating five digits in 60% of the cases, while 50% of those of the chronological ages 10 and 11 passed the test. The same group of subjects of the "mental age" of eight showed much more improvement in the test of copying a sentence, this test being passed by 35% of those age 8, by 65% age 9, by 94% age 10, and by 100% ages 11 and 12. The effect of maturity was much more marked in the second test.

The tests that showed the greatest increase with age were those of copying a sentence, writing from dictation, recalling a story read, and enumerating the days of the week. A slight increase was shown in the tests of playing the game of solitaire, knowing age, executing three commissions, counting backwards from 20 to 0, and showing the right hand. The increase was still less in the tests of repeating a sentence of 16 syllables and copying the diamond, and practically negligible in naming 5 coins. No increase was found in the tests of describing pictures, detecting omissions in pictures, counting 13 pennies, counting the fingers, comprehending easy problem questions, repeating five

digits, counting 9 "pfennig" (3 doubles and 3 singles), naming colors, comparing remembered objects, and defining in terms superior to use. According to Chotzen, the tests that show the greatest increase with age relate almost entirely to school training, those that show a slight increase depend to some extent on school training and other environmental experience, while those that show no increase with age involve largely factors of judgment and memory.

In treating his results according to the third method, Chotzen classified his subjects according to the final medical diagnosis he had made, as "not feeble-minded," morons (*Debilen*), imbeciles and idiots. The last group was dropped out, being small in number, and each of the other groups was divided into two parts according to age, one group being composed of subjects aged 8 and 9, and the other of subjects from 10 to 13. Certain tests stood out as differentiating very sharply the lines between the groups. For example, the test of comparing remembered objects was passed in the 8 and 9 year groups by 62% of the "not feeble-minded," 28% of the morons, and 5% of the imbeciles, while in the older groups it was passed by 93% of the "not feeble-minded," 69% of the morons and 10% of the imbeciles. The test of showing the right hand was passed in the younger group by 76% of the "not feeble-minded," 80% of the morons and 60% of the imbeciles. The first test showed a difference of 57% between "not feeble-minded" and imbecile children of 8 and 9, while the second test showed a difference of but 16% between these groups. The first test would therefore appear to be more diagnostic of intelligence.

The tests that showed the greatest diagnostic value in differentiating the groups of the younger children were copying the diamond, naming five coins, comprehending easy problem questions, comparing remembered objects and repeating five digits. The last two tests also differentiated the members of the older groups together with the tests of reproducing an item of a newspaper, arranging five weights, making change, defining in terms superior to use and naming the coins. Certain tests then are more val-

uable than others in the diagnosis of feeble-mindedness itself, and in diagnosing the different degrees of defect.

The results of the three methods of studying the individual tests agreed very closely on some tests. The tests of repeating five digits, comparing remembered objects and defining in terms of use, for example, showed a marked difference between normal and feeble-minded subjects, revealed no growth with age in children of approximately the same mental level and proved to be highly diagnostic of the different degrees of mental defect. In some cases the agreement was not as close. The test of remembering a story read showed a most decided increase with maturity independent of intelligence, but at the same time proved to be highly diagnostic of the mental status of "not feeble-minded," morons and imbeciles, and was one of the tests in which backwardness of the feeble-minded was most marked. The test of describing pictures, on the other hand, showed little difference between the performance of feeble-minded and normal subjects, and no growth with maturity. It can not be maintained then that the tests which show the greatest increase with maturity are those that are least dependent on intelligence, or conversely that those that show the highest correlation with intelligence are least dependent on maturity.

The lack of correlation between the results of the different methods is probably due to the faults of the methods themselves. The comparison of the results of feeble-minded children with those of normal children was made without any adequate statistical justification as to what percentage of a given age group must pass a test in order to have that test considered to be well within the ability of the group, or how large a difference there must be between the performance of feeble-minded and normal subjects in order to have that difference indicate more or less backwardness on that particular test. The personal equation of the investigator must necessarily play a large part in forming these judgments for the figures are quite irregular. The results of the second method are not conclusive, for Chotzen's figures do not

prove that the tests that show the greatest increase with maturity are least dependent on intelligence. The method demonstrates the influence of maturity and school training, but gives no information concerning the diagnostic value of the tests. The method of comparing the results of the different groups according to the final medical diagnosis was not entirely satisfactory, for the results themselves played some part at least in this diagnosis. If certain tests were given more weight than others, consciously or unconsciously, in making this diagnosis, then those tests would of necessity stand out as those that had the highest diagnostic value in differentiating the groups. This method depends to some extent at least on the personal equation of the investigator.

In general, however, although there was not a high correlation between the results of the different methods of study, certain tests were definitely shown to be dependent on factors of training and others were shown to be independent of this factor. Certain tests also stood out as diagnostic of feeble-mindedness while others were not as valuable in this respect. The investigation, though most suggestive, is not conclusive, and needs checking up and elaborating. The tests in the "ten," "eleven" and "twelve year" groups were not given enough to furnish any available data. As this is the region in which the scale breaks down, i.e. in diagnosing the higher grades of mental defect, it is most important that the work be elaborated here, and the relative diagnostic value of these tests determined.

The consensus of expert opinion shows that the Binet scale is not a reliable instrument for diagnosing the higher grades of feeble-mindedness. The problem before present investigators is the correction of this defect in the scale. The solution of the problem was indicated by Binet who, although he made all tests quantitatively equal, considered some tests more valuable than others in making a diagnosis. Descoeudres showed that the tests varied in the magnitude of their correlation with intelligence. Chotzen showed that all the tests were not of equal difficulty for feeble-minded children. The Binet scale has been shown to be

composed of some tests that are effective in diagnosing intelligence, and others that are ineffective. To make the scale more effective it is necessary to determine the relative diagnostic value of the individual tests.

During the process of analysing the Princeton data, the writer compared the results of 49 children who had been rated "dull" by the teachers and 46 children of the same age who had been rated "bright." The results were striking. Some tests were as easy for the dull children as for the bright children. Other tests were easy for the bright children but practically impossible for the dull children. The results of this study led the writer to undertake a much more extensive investigation in the Trenton, N. J. public schools, where on account of the large number of children in the schools it was possible to obtain two groups of known differences of intelligence. To these groups the Binet and other tests were given in order to determine what sorts of tests were most highly diagnostic of intelligence. The results of this investigation are contained in the following chapters.

II. CHARACTER OF SUBJECTS

Two methods are open to the investigator who wishes to determine the relative diagnostic value of mental tests. The first is that of giving the tests to a group of individuals who can be independently and accurately rated in their rank order of intelligence. From the standing of the individuals in the various tests, the correlation of the tests with intelligence may be obtained. The other method is that of comparing the results of two or more groups of known differences of intelligence. The differences in the performances of the groups on the tests will indicate the diagnostic value of the tests. The second method was followed by the writer as it was thereby possible to obtain a more objective indication of the intelligence of the children examined.

In speaking of intelligence, the writer means by the term exactly what Stern (62) meant: "general mental adaptability to new problems and conditions of life" (page 3). The objective indication of intelligence that was used was that of school standing. The public school presents a situation that must be met by all children. Some children meet the situation adequately and pass through the schedule of grades in the required time. Others are unable to adapt themselves to the changing conditions and drop farther and farther behind their fellows. A retardation in school of one or possibly two years may be due to a great many causes independent of the mental status of the individual, but a retardation of more than two years would probably indicate, in the absence of a more obvious explanation, an inferior intellectual endowment.

The experiments reported here were conducted on two groups of children, one group composed of boys in the regular grades of the Franklin school in Trenton, N. J., the other group of members of the special classes for backward and defective children or candidates for admission to these classes in the same city. The

law of New Jersey makes it mandatory that children showing three or more years pedagogical retardation be placed in special classes. The pedagogical retardation that is meant is that indicated by the age of the individual and the grade he should be in at that age. Inasmuch as the school grade age used in this way does not take account of late entrance or enforced absence, the criterion of retardation used in this investigation was that of grade progress or in other words the relation between the grade of the individual and the number of years he had been in school.

In order to find a group of retarded children the writer examined nearly all of the boys and girls in the Trenton schools who were either in special classes or who were candidates for admission to these classes. 229 boys were examined, 203 being in the classes while the remaining 26 were candidates for admission to the classes. All of these subjects were given the Binet tests, and those whose retardation showed no obvious explanation other than inferior intellectual endowment were given the complete series of tests.

Concerning each subject examined, the following information was obtained:

1. AGE. Obtained from school record and checked up by questioning subject. In cases of doubt, the families were consulted or the age was based on certificates of birth or baptism.

2. NUMBER OF YEARS IN SCHOOL. Obtained by the teacher and checked up from school records and by questioning subject. The actual number of years in school was used, exclusive of absence on account of long sicknesses, etc. Time spent in parochial schools where foreign languages were taught was not counted as time in school, so that the number of years in school actually means the number of years in English speaking schools.

3. GRADE. Estimated by teacher in case of children in the ungraded special classes. The teacher estimated the grade that the pupil could enter if he were to be transferred to the regular grades. The basis of the teacher's judgment was the work of the pupil in the reading books, spelling books, arithmetic books, etc., standard for the different grades. All of the teachers had taught in the regular grades, and all of the pupils at one time

or another had been in the regular grades. The teacher's estimate was checked up from previous estimates independently made, and from the monthly and annual reports of the pupil.

4. NATIONALITY, LANGUAGE USED AT HOME, PLACE OF BIRTH. Obtained from teacher and checked up by questioning subject.

5. PHYSICAL DEFECTS. Obtained from record of medical examination.

93 boys in the special classes were given the complete examination. The subjects ranged in age from 9 to 16 with the majority at the ages 12, 13 and 14. This latter group of 59 boys 12, 13 and 14 years of age is used for comparison with a group of 58 boys of the same age in the regular grades. The first group will be spoken of as the retarded group, the second as the normal group. 4 members of the retarded group were candidates for admission to the special classes, the remainder being regularly enrolled in these classes.

The character of the normal and retarded groups may be studied by comparison with the general run of all children in school. The members of the two groups had been in school 5, 6, 7, 8 or 9 years. The grade progress distribution of all children in the Trenton schools who had been in school 5, 6, 7, 8 or 9 years is shown in Table I which is derived from the "Grade and Progress" table on page 104 of the Report of the Trenton Board of Education for 1914.

TABLE I.
Grade Progress Distribution of 4323 Trenton Children in School 5, 6, 7, 8 and 9 years.

GRADE	Number of Years in School.					Totals
	5	6	7	8	9	
I	3	1				4
II	25	10	4			39
III	175	45	5	4		232
IV	460	181	61	6	1	709
V	586	376	130	28	9	1129
VI	157	479	272	62	4	974
VII	22	108	374	169	37	710
VIII	5	25	145	233	118	526
Totals	1433	1225	994	502	169	4323

By inspection of Table 1 it may be seen that the largest proportion of the 4323 children are "at grade" or in the grade corresponding to the number of years that they have been in school, i.e. the fifth grade for those who have been in school five years, the sixth grade for those in school six years, etc. The percentage distribution of the children "at grade" and at each year above and below grade is as follows:

+3	+2	+1	"at grade"	-1	-2	-3	-4	-5
0.1%	1.1%	9.5%	38.8%	32.2%	13.5%	3.8%	0.8%	0.2%

This distribution is shown graphically in the top portion of Fig. 1. By far the largest proportion (80.5%) are "at grade" or one year above or below grade, the range of the distribution of the 58 normal subjects. 18.3% are two or more years retarded, the range of the distribution of the 59 retarded subjects. Only 1.2%

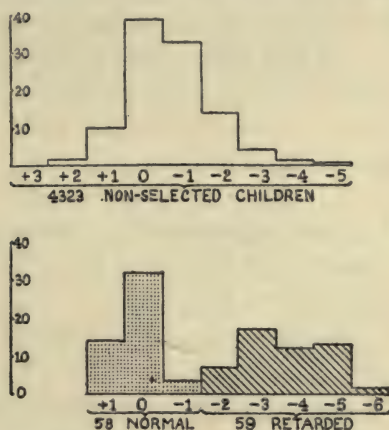


FIG. 1. Grade Progress Distribution of Normal and Retarded Subjects and of all Children in School 5, 6, 7, 8, or 9 years.

are two or more years ahead of their grade. This distribution can not be taken as typical of the entire school course, for there are no grades above the eighth included, so that there are no above grade groups possible for children 8 or 9 years in school, and no "at grade" group for those in school 9 years. The distribution has value only for comparison with the children actually tested. The grade progress distribution of the 117 subjects whose data are used in the subsequent report is shown in Table 2.

TABLE 2.
Grade Progress Distribution of 117 Subjects of This Investigation.

Grade	Years in School					Totals
	5	6	7	8	9	
I		1				1
II	1		3	1		5
III	1	3	4	5	1	14
IV		4	11	8	6	29
V			3	5	2	10
VI		2				2
VII		5	21	1		27
VIII			11	14	4	29
Totals	2	15	53	34	13	117

It may be seen by inspection that Table 2 is composed of two characteristically different groups, 59 boys who have been in school from 5 to 9 years and are in Grades I, II, III, IV, and V, and 58 boys who have been in school the same length of time but are in grades VI, VII and VIII. The percentage distribution of the children "at grade" and at each year above and below grade is as follows:

+ 1	"at grade"	- 1	- 2	- 3	- 4	- 5	- 6
13.7%	31.6%	4.3%	6.8%	17.1%	12.0%	12.8%	1.7%

This distribution is shown graphically in the lower part of Fig. 1, in which the dotted portion represents the normal group and the shaded portion the retarded group.

Comparing the distribution of the subjects tested with that of the general run of children, the range of the normal group, i.e. "at grade" or one year above or below grade, corresponds with 80.5% of the 4323 cases. The range of the retarded group corresponds with that of the lowest 18.3% of these 4323 cases. The two distributions are actually farther apart, for they represent the extreme samples of the children tending toward normal or accelerated grade progress and toward retardation. The largest proportion (86%) of the retarded group are 3 or more years retarded or within the range of the lowest 26% of all the 794 children two or more years retarded. The largest proportion (92%) of the normal group are either "at grade" or one year ahead of grade, a range falling within the highest 61% of the 3529 children one year retarded, "at grade" or accelerated.

The same facts concerning the extreme divergence of the two groups examined are brought out by a study of the ages in grades. From the "Age in Grade" table on page 113 of the Report of the Trenton Board of Education for 1914, the grade distribution of all 12, 13 and 14 year boys was obtained. From these data the average grade was computed for all 12, 13 and 14 year boys. The average grade of the normal and retarded groups was also computed at these ages. These averages are shown in Table 3.

TABLE 3.

Average Grade of 12, 13 and 14 year Boys.

Age	All Trenton Boys		Normal Group		Retarded Group	
	No.	Ave. Grade (MV)	No.	Ave. Grade (MV)	No.	Ave. Grade (MV)
12	636	5.21 (1.57)	18	6.94 (0.15)	18	3.22 (0.91)
13	644	5.66 (1.92)	20	7.50 (0.50)	21	3.90 (0.43)
14	354	6.85 (2.05)	20	7.90 (0.18)	20	3.95 (0.58)
all subjects			58	7.47 (0.53)	59	3.71 (0.72)

The average number of years that the retarded group have been in school is 7.42 yrs. (MV=0.83 yrs.), while the average number for the normal group is 7.28 yrs. (MV=0.61 yrs.). In spite of the fact that both groups have been in school the same length of time, the normal group averages over three and a half grades higher than the retarded group. The retarded group have gone just half as far as the normal group in the same length of time.

The age in grade distribution of the 117 subjects 12, 13 and 14 years of age is shown in Table 4.

TABLE 4.

Age in Grade Distribution of 117 Subjects.

GRADE	12	13	14	Total
I	1			1
II	4		1	5
III	5	5	4	14
IV	6	13	10	29
V	2	3	5	10
VI	2			2
VII	15	10	2	27
VIII	1	10	18	29
Total	36	41	40	117

Although the retarded subjects have been in school as long as the normal subjects, none of them are above the fifth grade. The largest proportion of the normal group (96.5%) are in the seventh and eighth grades.

Such then are the differences between the two groups. It will be seen subsequently that when tests are given to the two groups, some of the tests are equally easy or equally difficult for both groups, while others differentiate the groups sharply. In order to consider these tests as diagnostic of intelligence it is necessary to show that a real intelligence difference exists between the two groups. The writer's thesis is that the gross differences in school progress of the two groups indicate a real difference in intelligence. The retarded group although in school as long as the normal group have made only half the progress. The average retardation of three years and a half indicates an inferior intellectual endowment. To show this it is only necessary to prove that this retardation is not due to any other cause.

The most frequently mentioned cause of retardation other than intellectual defect is that of illness, which causes irregularity of attendance and consequent failure to keep up with the class-room work. As far as the writer was able to learn from the medical records, none of the retarded group had had any serious illness which kept him out of school any considerable length of time. Nor were there any uncorrected physical defects. The special classes contained children with physical defects such as very poor vision or deafness, but none of these cases was included in the retarded group. All of the special class children had annual eye examinations and the defects were corrected. As far as could be ascertained the members of the retarded group were physically normal.

Another frequently mentioned cause of retardation is an educational curriculum which is maladapted to the needs of the individual. Further than this it is claimed that the teaching is inferior and the lack of individual instruction a handicap. Faults in the educational regime would cause retardation in both groups of subjects, for both groups were exposed to the same regime. These causes of retardation would not explain why one group was retarded and the other not retarded.

Deficiency in the use of language is frequently given as the cause of the retardation of the children of non-English speaking parents. Since the retarded group contains 20 children of non-English speaking parents, it is possible that these children were retarded on account of language deficiency. The average number of years that these 20 children had been in school is 7.7 yrs. ($MV=0.73$ yrs.). The average grade of the group is 3.7 ($MV=0.72$). If this retardation of four years is to be attributed to language deficiency, it is legitimate to expect that the same factor would influence the normal group. The normal group contained 22 children of non-English speaking parents. The average number of years in school of this group is 7.36 yrs. ($MV=0.65$ yrs.), and the average grade 7.58 ($MV=0.52$). Instead of showing a retardation these boys show a slight advance. Although they have been in school the same length of time as the retarded group, they have progressed twice as far. It is not possible then to account for the differences between these two groups on the basis of language deficiency. Both groups had the same language handicap, and the difference in school progress must be referred to the inferior intellectual endowment of the members of the retarded group. The reactions of retarded and normal, English and non-English groups on the individual tests will be studied later to determine the influence of language training on the tests.

It is legitimate to conclude that the gross differences between the school progress of the retarded and normal groups indicate a difference in intellectual endowment. The members of the retarded group are retarded because they are less intelligent. The writer would not take an extreme position, and say that in all cases retardation indicates intellectual deficiency, or that presence in the proper grade indicates intellectual normality. He is willing to admit that a child may become seriously retarded by a particularly unfortunate combination of circumstances or by a lack of push or interest quite independent of his intellectual make-up. These cases are exceptional however. On the other hand, the writer would be willing to admit that subnormal children by a combination of fortunate circumstances may be pushed

ahead, and may find themselves in a grade altogether beyond their ability. The writer is certain that the retarded group contains some perfectly normal boys, and that the normal group contains some members that are intellectually inferior to some members of the retarded group. These cases are exceptional however, and are far from the general run. On the whole, the retardation of the backward group indicates inferior intellectual endowment. Tests which sharply differentiate the normal and retarded group are those which are diagnostic of intelligence.

This position may be extreme, and it is impossible to justify it on any other grounds than personal opinion. If the school measure were an adequate measure of intelligence there would of course be no need of mental tests. No one who has had experience in diagnosing mentality will ignore the school record entirely, nor will he stress it above all other measures. The distribution of ability given by the school measure is most certainly not the distribution of general ability, for the school gives practically no opportunity for intellectual precocity to show itself. The lack of confidence that experienced persons show in accepting the school record as an absolute indication of intellectual inferiority is also evidence that retardation does not always indicate intellectual backwardness. The writer believes however that on the whole there are certain large intellectual differences between the normal and retarded groups. It was not possible to account for the presence of any member of this group in the special classes for backward and defective children on any other grounds than mental defect, and the histories of the individuals were very carefully studied. With the same physical and environmental opportunities they have made half the progress of their brothers in the regular grades. The tests which differentiate the members of the retarded and normal groups are at least diagnostic of pedagogical retardation. On the basis of the extreme pedagogical retardation, the writer believes such tests to be diagnostic of intelligence.

It is not necessary in this study to classify the retarded group in terms of the number of normal individuals, morons and imbeciles. The giving of diagnoses would involve the presentation

of case histories that would require considerable space. The definition of the group in terms of school progress is sufficient. The retarded group by no means represents the lowest selection of the general population, for idiots do not enter school, and the low grade imbeciles are usually removed from school before they are 12, the lower age limit of the retarded group. A few cases were examined (not more than five) who presented the proper qualifications of age, parentage and scholastic standing to be included in the retarded group, but who were mentally so low that it was useless to put them through the series of advanced tests used. The selection of retarded subjects is therefore rather high. No member of the retarded group tested below the "mental age" of eight. The distribution of ability in the retarded group may be fairly said to range from that of the low grade moron to that of the normal individual. The group was composed mostly of border-line cases of feeble-mindedness. The lowest members of the group probably correspond to the highest grade of our present institutional cases. The group is therefore singularly well adapted for comparison with a normal group to determine what tests are most efficient in detecting the higher grades of mental defect.

III. TESTS AND PROCEDURE

The arrangement of tests used was that of Binet's 1911 scale (8). Town's (72) translation was used, but not followed strictly throughout. The translations of some of the tests were taken over from Goddard (28), as the freer translation seemed better adapted to the linguistic training of the subjects. The instructions given were partly those of Binet, partly those of Goddard, and partly those of the writer. Inasmuch as there were many departures from the standard form of procedure, the procedure used is given in detail below. The close student of standardized procedure will probably find several startling heresies in the account. In applying the tests, however, the writer was not seeking to find the "mental age" of the subject, or to obtain age norms for the tests. He was primarily interested in testing the tests by comparing the results of normal and retarded children. In order to discover the factors involved in each test, it was very necessary that the subject should understand the nature of the task he was to perform. For this reason a fore-exercise or practise test was given wherever possible. The important factors in the experiment were to be certain first, that the subject understood the nature of the task, and second, that the instructions were uniform for both groups of subjects.

The detailed account of the instructions illustrates the method of providing for the first factor. The retarded group was examined first so that the procedure is largely adapted to their needs. Although all of the Binet tests were used, only those are reported that were given to all the selected retarded and normal subjects. The list of Binet tests used, the procedure adopted, and the criteria for scoring each follow:

AGE III

No. 3. ENUMERATING OBJECTS IN PICTURES. Binet's original uncolored pictures were used. The question asked,—“Tell me what you see in this picture?” The question

was repeated for each picture. The picture of "A man and a boy" was shown first, of "A man and a lady," second, and "A prisoner," third. The responses were scored according to the instructions in the categories of enumeration, description, interpretation and emotion. No time limit.

No. 5. REPETITION OF SENTENCES. Instructions,—
"I am going to read you a sentence and I want to see how well you can remember it. Say it just as I say it. Don't leave out any words." The following list of sentences was used:

10—My name is William. Oh the naughty dog.

12—It rains in the garden. John has finished his task.

14—We are enjoying ourselves greatly. I have caught a mouse.

16—Let us go for a long walk. Give me the pretty, little bonnet.

18—Mary has just torn her new dress. I have given two cents to that beggar.

20—It is not necessary to hurt the birds. It is night, all the world rests in sleep.

22—We expect to have a great time at the seashore, digging in the white, beach sand all day long.

24—My little children you must work very hard for your living, you must go to school every day.

26—The other day I saw on the street a pretty, yellow dog. Little Maurice has stained his nice new apron.

28—Ernest is frequently punished for his bad conduct. I bought at the store a pretty doll for my little sister.

30—There was a severe storm last night with much lightning. My comrade caught cold, and he now has a high fever, and coughs a great deal.

32—The car is less expensive than the omnibus, it costs but two cents. It is strange to see women acting as coachmen in Paris.

All of the sentences were taken from Town's translation except No. 22, which was taken from Whipple's Manual (75) pg. 494. The word "Mary" was substituted for "Charlotte" in sentence No. 18, and in No. 24, the word "day" was used instead of

"morning," in order to make 24 syllables. The subject was scored according to the number of syllables in the longest sentence that he succeeded in repeating without error. After arriving at this "threshold," the subject was given the next two longer sentences, to make sure that he had reached his limit. Near the close of the experiment, this procedure was found to be entirely inadequate, so that the results from this test are unreliable. No time limit.

AGE V.

No. 3. See III, 5.

AGE VI.

No. 2. DEFINING IN TERMS OF USE. The words used were "fork," "table," "chair," "horse" and "mother," being given in the order named. The question asked was "What's a fork?" or "Tell me what a fork is." The responses were scored as definitions by use or superior to use according to the criteria given in the previous discussion of the personal equation (see page 19). No time limit.

AGE VII.

No. 2. DESCRIBING PICTURES. See III, 3.

AGE VIII.

No. 4. GIVING THE DAY AND DATE. An error of four days was allowed on either side of the day of the month. Scores were recorded for each part of the test, the day of the week, the month, the day of the month and the year. No time limit.

No. 5. REPEATING 5 DIGITS. "Say these figures just as I say them." 47395, 51742, 83964. One out of three scored. The experimenter's rate of giving the digits was somewhat faster than two per second, but was constant through habit.

AGE IX.

No. 1. MAKING CHANGE. "How much is six cents from a quarter?" The coins were not given. No time limit.

No. 2. DEFINING IN TERMS SUPERIOR TO USE. See VI, 2.

No. 3. NAMING PIECES OF MONEY. "Tell me all of the pieces of money you know." If the subject stopped at 50c, he was asked "What comes next?" In naming the bills, if he said "Three dollars, four dollars" etc., he was asked if he had ever seen a three dollar bill. None of the coins or bills were shown. No time limit.

No. 4. NAMING THE MONTHS. "Say the months for me." One error of omission or inversion was allowed. Time allowed, 15 seconds.

No. 5. COMPREHENDING EASY QUESTIONS. The following questions were asked verbatim:

- a. What would you do if you missed a train?
- b. What would you do if you had been struck by a playmate who didn't mean to do it?
- c. What would you do if you had broken something that didn't belong to you?

The order of presentation was *a, c, b*. No time limit.

AGE X.

No. 1. ARRANGING 5 WEIGHTS. Stoelting's standard cubes were used. Three trials were given and the number of successful attempts recorded. As a control on this test, a series of definitely supraliminal weights was introduced. The weights were made of metal salve boxes, and were weighted at 20, 30, 45, 70 and 100 gms, roughly in accordance with Weber's law, so that the subjective differences between the weights was approximately equal. One trial with these weights was given before the standard weights. If the subject failed, the failure was obviously due to the intellectual inability to comprehend a serial arrangement, or to make the logically necessary comparisons. If he passed, the test served as a "warming up" or practice test. Subsequent failure on the Binet weights would obviously be due to failure in sensory discrimination. The instructions were "Put these boxes in a row with the heaviest one first, and then the next to heaviest, and then the next, and then the next and then

the next." The experimenter pointed at a spot on the table for each box. The weights were not touched, and no suggestion was given for the subject to lift them. No time limit.

No. 2. COPYING DESIGNS FROM MEMORY. "I am going to show you this card with two drawings on it: I'm only going to show it to you for a little while, so I want you to study it hard and remember what it looks like." While these instructions were being given, the card was turned over as rapidly as possible, but in such a way that the subject caught a glimpse of it and understood the test better. The reproductions were scored at the time of the experiment, and then independently re-scored according to the arbitrary point system outlined in the previous discussion of the personal equation. The duration of the exposure was 10 seconds.

No. 3. DETECTING ABSURDITIES IN STATEMENTS. "I am going to give you a sentence that's got something foolish in it, and I want you to tell me what's foolish.—Now what's foolish in that?" Absurdity *d* was given first. If the subject failed to see the point, it was explained to him. After the first one, no explanations. The order of presentation was *d*, *c*, *a*, *b*, *e*, each test being separately scored. The following absurdities were given, the same wording being adhered to throughout. No time limit.

a. An unlucky bicycle rider fell off his bike and broke his neck. They took him to the hospital, and they don't think he'll get well.

b. I have three brothers, Henry, Robert and myself.

c. Yesterday, there was a railroad accident, but it wasn't a bad one, only 48 people were killed.

d. Yesterday, the police found a body of a girl cut up into 18 pieces. They think that she killed herself.

e. A man once said, "If I were going to kill myself, I wouldn't do it on Friday, for Friday is an unlucky day, and it might bring me bad luck."

No. 4. COMPREHENDING DIFFICULT QUESTIONS. The order of presentation used was *a*, *b*, *d*, *e*, *c*. The following questions were given, the same wording being adhered to throughout. No time limit.

- a. What would you do if you were delayed in going to school?
- b. What would you do before taking part in an important affair?
- c. Why is it easier to forgive a bad action done in anger than it is the same action done without anger?
- d. What would you do if you were asked your opinion about someone you didn't know very well?
- e. Why should you judge a person by his actions rather than by his words?

No. 5. USING THREE GIVEN WORDS IN A SENTENCE CONTAINING TWO IDEAS. A practice test was given as follows,—“I want you to make me up a sentence, a good sentence, with the words ‘boy, play and sled’ in it.” After the subject had given the sentence and had shown that he understood what was expected of him, the words “Trenton, money and river” were given. The subject gave the sentence orally. No time limit.

AGE XII.

No. 1. RESISTING SUGGESTION. (LENGTH OF LINES). The lines were accurately drawn by a draughtsman, on Bristol board 5 by 28 cm. The question asked for each of the first three cards was “Which is the longer line here?”, and for each of the last three, “And here?” The number of correct responses was noted.

No. 2. USING THREE GIVEN WORDS IN A SENTENCE CONTAINING ONE IDEA. See X, 5.

No. 3. GIVING 60 WORDS IN THREE MINUTES. “I want you to give me all the words you can think of in three minutes. Any old word will do, like ‘man-beard-boy-shirt-carriage,’ now go ahead.” The actual number of words was recorded.

No. 4. DEFINING ABSTRACT TERMS. “Charity, justice and kindness” were used. The question asked was “What’s charity?” or “Tell me what charity is?” A concrete illustration of an act was considered acceptable. No time limit.

No. 5. RECONSTRUCTING DISSECTED SENTENCES. “I am going to show you a card with a sentence on it that’s all mixed up. The words are in the wrong order, and I want you

to tell me what the sentence would be if it were in the right order." The following sentences were used, presented in the order *b*, *c*, *a*. One minute was allowed for each.

- a.* started-the-for-an-early-hour-we-country-at
- b.* a-défends-dog-good-his-master-bravely
- c.* asked-paper-the-to-I-teacher-correct-my

AGE XV.

No. 1. REPEATING 7 DIGITS. 2964375, 9285164, 1395847.

No. 2. GIVING THREE WORDS RHYMING WITH DEFENDER. As an introductory test, the subjects were asked to give rhymes with "day, mill and spring." An example was given, "man, ran, can, fan" and then the subject asked to give rhymes with "cat." If he failed, rhymes with "cat" were suggested until he either succeeded or failed utterly to comprehend. If he succeeded, he was then given the stimulus word in the order D, M, S, and allowed half a minute on each. At the end of this list, the stimulus word "defender" was given, and the subject allowed one minute.

No. 3. REPEATING A SENTENCE OF 26 SYLLABLES. See III, 5.

No. 4. INTERPRETING PICTURES. See III, 3.

No. 5. SOLVING PROBLEMS FROM VARIOUS FACTS. "I am going to tell you a story and I want you to tell me what's happened." The following wording was used in the problems:

a. A lady was walking through a park out in Chicago, when she suddenly stopped, very much frightened, turned around and ran back to the nearest police station, and there she told them that she had seen, hanging from the limb of a tree—What did she see?"

b. My neighbor has been having strange visitors lately. One after the other, a doctor, a lawyer and a priest (minister) have called at his house. What's happened there?"

The word "minister" was substituted for Protestant children. The only answer accepted to the first question was "a man hanged," to the second question, "someone's very sick," "just

died" or "is dead." The subject was cross-examined to make sure he had taken all the facts into consideration. If not, the test was failed. Question *a* was given first, *b* second. No time limit.

The method used by the experimenter in obtaining uniform procedure was that of practice. Before undertaking the Trenton experiment, the writer had examined over 150 children with Goddard's 1911 scale, that experience serving as an adequate practice series at least. The accuracy and uniformity of technique in giving the Binet tests rests entirely on the previous experience of the experimenter. The Binet tests used were, however, only a small portion of the complete examination given to the retarded and normal groups. The greatest difficulty was experienced in obtaining uniform procedure in the tests other than those in the Binet series. The method of obtaining uniformity of procedure in the supplementary tests is outlined in a subsequent section devoted to the discussion of these tests.

The complete series of tests was very long, and contained many tests such as that of repeating sentences that were not attractive to the subjects, so that great care had to be used by the experimenter in arranging the series so as to maintain the subject's interest. The complete series included besides the Binet tests described several special tests of memory, suggestibility, discrimination, reasoning ability and a long series of puzzle tests. The entire list of tests was given to only 153 of the 289 subjects examined.

The tests that were interesting depended largely on the subject. Any test that was obviously too difficult for a subject was uninteresting to him. The converse is usually true in practice also, that a test that is well within the subject's range of ability is interesting to him. The free play of thought on a novel situation is in itself pleasant if that situation offers a ready solution—the unpleasant tone arises with the attitude of doubt. For this reason the factor of cooperation in the mental examination of normal individuals is not as difficult as persons who have not attempted it would have us believe. The factor of cooperation in the examination of pure defective types is by no means an insurmountable difficulty, if the examiner has tact and discretion.

It is only in cases of mental alienation that the factor occasionally becomes an insurmountable difficulty, and none of the subjects in the finally selected groups presented any developed aberrational tendency.

When an individual is examined by the incomplete method of testing, the experimenter can keep the questions well within the subject's range of ability, and by a judicious selection of the questions can maintain his co-operation without difficulty. In this experiment, however, all the questions were given to every subject, and it was necessary in the case of most of the retarded group to hammer them through a long series of questions that were obviously impossible for them to answer, so that the experimenting was exceedingly difficult.

The duration of the complete examination was from an hour and a half to two hours depending on the individual. The experiment was always broken up into periods usually of half an hour each. No subject was ever kept over 45 minutes, and only that long in exceptional cases. The most difficult questions were given at the beginning of the period. If the subject showed signs of fatigue, he was immediately dismissed. Although the character and duration of the examination was such that the factor of cooperation would appear to be very large, the writer found little difficulty from this source. The special classes were composed of about 15 children. The plan of attack in beginning a new class was to have the teacher send the most popular boy in the room first, and the experimenter would do his best to show him a good time. The puzzle series and similar tests proved an unfailing source of pleasure. The first examination of the most popular boy contained the easiest and most interesting tests of the series, and the decoy invariably worked. In every class except one, the teacher would send the boys to the examiner as a reward for good conduct in the class room. The experimenter in turn saved the puzzle tests and easy tests as a reward for careful work in the more difficult tests, always saving some of the attractive tests for the close of the examination so that the subject would return to his room in good spirits. The examinations were conducted for the most part during the school hours, and it was found that the subjects were glad enough to avoid their lessons

for a more attractive occupation. The examination was never allowed to conflict with the gymnasium hours. When an examination was conducted outside of school hours, the subjects were paid at the rate of twenty cents an hour. If the subjects showed the slightest lack of interest or the slightest unwillingness to return for a second examination, the experiment was abandoned. In one class after three or four boys had been successfully examined, one boy returned with an unfavorable verdict, the rest of the pupils showed a slight unwillingness to come, and no further experimenting was done. On the whole, the greatest difficulty was experienced not in enticing the boys to come, but in arranging waiting lists and keeping the boys out of the room when they were not wanted. Very little difficulty was experienced from the factor of cooperation.

The conditions under which the examinations were made were not ideal, but were uniformly satisfactory in every case. A separate room was always provided. A table was arranged so that the light coming from behind the subject fell directly on it. In a few cases, the experimenter could not prevent the presence of a third person in the room. If this proved a distraction, the examination was discontinued, or if the person wished to observe the test, the examination was completed and the record discarded.

The factor of information was encountered, but the nature of the tests was such that information played little part. The examination was so varied, and so long, that the subject would have difficulty in remembering any but a few striking tests. The test of naming 60 words gave the most difficulty on this score, the subjects invariably telling the experimenter that they were "all loaded" for this test. The special classes were in various schools throughout the city, but the normal pupils were all examined in one school, so that the factor of information would be more in their favor. As a matter of experience, however, the experimenter had no difficulty in finding out whether or not the subject had any previous information. The record on any test was discarded if this factor were discovered. The results of the 60 word test are given, however, with this reservation. In any test in which the presence of this factor was suspected, the writer will report it in the discussion of that test.

IV. VARIABLE FACTORS

When the results of the normal and retarded groups are compared, it is seen that some tests are equally easy or equally difficult for both groups, while other tests differentiate the groups sharply. In order that the latter type of test may be considered diagnostic of intelligence, it is necessary to show that the difference in the reaction of the two groups may not be attributed to any other factor than intelligence. In the preceding study, several variable factors were found which might influence the outcome of the tests, and these factors will be discussed in conjunction with those of sociological conditions and language training. The variable factor of sex differences of course drops out, since only boys were included in this investigation.

THE ERROR DUE TO INCOMPLETE DATA.

The error due to incomplete data can have no effect on the results, for practically all the tests were given to all the subjects of both groups. From time to time a test would be dropped out because it was incorrectly given, or a test would be accidentally omitted, but on the whole the tests were given very completely. Taking the Binet tests, it was possible to give to the retarded group 3363 tests. Of this number, 3314 tests or 98.5% were given. The normal group were given 3280 tests or 99% of the 3306 tests possible to be given. Question *a* of the comprehension questions was given to the normal group only 74% of the possible number of times. The date test was given to this group 95% of the time, and all the other tests were given over 98% of the possible number of times. The retarded group were given the date test 86% of the time, parts *d* and *e* of the definitions test 92% and 88% of the time, and all other tests over 97% of the possible number of times. The testing may be called complete, then, and the results cannot be influenced by the error due to incomplete data.

THE INFLUENCE OF GRADE TRAINING.

The results of the tests may be influenced by grade training, but there is no way to determine the influence of this factor. The only difference between the training of the children in the special classes and in the regular grades, consists in the larger amount of time devoted to manual training in the special classes. Very nearly a fifth of the time is given to work of this nature. The rest of the time is devoted to regular school work, and the only difference between the special classes and the regular grades in this work is that the former receive individual help and instruction almost entirely. They have more actual training in the school subjects than the children in the regular grades but they do only half as well. The difference in school standing between the normal and retarded groups must be attributed to a difference in intellectual endowment, for both groups had had an equal amount of training along the same lines. Regardless of how they took the training, the retarded group had been exposed to it as long as the normal group.

THE INFLUENCE OF THE PERSONAL EQUATION

The differences between the two groups may be due to the influence of the experimenter's personal equation, but there is very little check on this factor due to the fact that there was only one experimenter, the writer. The test of copying the designs from memory was ranked plus or minus by the writer at the time of the examination, and later scored according to the arbitrary point system described in Chapter III of the first study (see pages 23 to 26). It is possible to compare the experimenter's judgments at the time of the examination with the later arbitrary scoring which was made without knowledge of the original rank given. According to the experimenter's first judgments, 69% of the normal group and 41% of the retarded group passed this test. According to the point system of scoring in which 15 points is used as the passing mark, 67% of the normal group and 31% of the retarded group pass the test. The score of the normal group is reduced 2% by correction, the score of the retarded

group, 10%. The passing mark for each group was calculated as in Chapter III of the first study, and was found to be 14 points for the normal group and 13 points for the retarded group. The experimenter marked as failed in the normal group one design scoring over 15 points, and marked as passed two designs scoring under 15 points. In the retarded group, on the other hand, the experimenter marked as failed two designs scoring over 15 points, and marked as passed nine designs scoring under 15 points. The experimenter was therefore more lenient with retarded than with normal children in the test of copying designs from memory.

It is not possible to obtain a quantitative estimate of the influence of the personal equation in any of the other tests. It is the opinion of the writer that if the personal equation influenced the results at all, the retarded group were favored. The experimenter examined 266 children in the special classes before examining the normal children. The work of the special class children was uniformly so low that comparatively poor answers to some questions would seem to be very good. The tests of detecting absurdities in statements and comprehending difficult problem questions were rarely answered in the special classes. An answer that would have been considered doubtful if given by a normal subject, would probably have been ranked plus if given by a member of the retarded group, owing to the contrast of his answer with those of the other retarded children. As far as possible, the experimenter used a uniform system of testing and scoring, but the criteria for judging some tests are at best indefinite, and are therefore susceptible to the recent experience of the experimenter. The writer is not conscious of favoring the retarded group, but this tendency might have been present. Considerable tact and patience was demanded in giving the difficult tests to subnormals, and the writer is certain that if the tests had been given to the special class children after the subjects in the regular grades had been examined, he would have been much more abrupt and less patient.

The fact that there was only one experimenter would tend to minimize the influence of the personal equation as it would make

impossible any large differences of procedure or technique. In spite of this fact however it is possible for this influence to be present. The Binet tests are scored by the all or none method, they are ranked either plus or minus, while the responses on many of the tests are not all or none responses. The 60 word test, for instance, is ranked plus if the subject gives 60 words or over, and minus if he gives under 60 words. The responses of the subjects of this investigation actually varied from 31 to 196 words. The degree of merit on this test may vary through 165 steps by steps of 1, while the expression of merit used is merely plus or minus.

The tests vary in the degree in which the responses may be accurately rated by the all or none scoring. Some tests, such as that of making change, have an all or none response—they are either right or wrong. Other tests may admit of a slight grading of response. The 5 weight test has four grades of response according as the subject arranges the weights correctly 0, 1, 2, or 3 times. At the other extreme are tests such as the design test that admit of at least 20 grades of response. Some tests would admit of considerable grading, but there is no available method for grading them. The responses to the test of using three given words in a sentence containing one idea, for instance, would vary from "Trenton's river costs money" and "Trenton has lots of money on the river" to "Trenton paid a large sum of money to have the Delaware river deepened" and "The people in Trenton that have money live along the river." It is conceivable that the responses to this test, and to other tests such as defining in terms superior to use, defining abstract terms, and some of the absurdity and comprehension questions could be arranged on a scale of merit from 0 to 10.

Given a scale of merit from 0 to 10, the experimenter must express his judgment by plus or minus. At some point on this scale there is bound to be a range of uncertainty—a range of play for the personal equation. In the preceding investigation it was found that Experimenter C changed his criteria in ranking the definitions test during the course of the experiment (see

page 20), so that it is possible for the personal equation to enter even with one experimenter. If the experimenter's judgment changed in examining normal and retarded subjects, the diagnostic value of the tests would be influenced. If the experimenter were too lenient with retarded and too severe with normal children, the diagnostic value would be less than it should be. The design test showed this influence, and the diagnostic value was lowered 8% by a variation in judgment of 1 point on a scale of merit of 20 points. If the opposite tendency were present—if the experimenter were too severe with retarded and too lenient with normal subjects, the diagnostic value of the test would be exaggerated. It is possible that this should happen, for a range of uncertainty is present, and the influence of the personal equation too subtle to be eliminated entirely.

It is possible then for a person to maintain that the differences in the reactions of the groups which are used to determine the diagnostic value of the tests are merely expressions of the personal equation of the experimenter. The safest position to take probably lies between the two extremes—to bear in mind that the diagnostic value may be thrown one way or another by the personal equation, but also to bear in mind that more of the responses are bound to fall in the range of certainty than in the range of uncertainty, and that therefore the personal equation can not vitiate the results entirely.

THE INFLUENCE OF SOCIOLOGICAL STATUS.

In order that the differences between the performance of the retarded and normal groups on the tests might not be referred to environmental conditions and home training, the writer was very careful to select subjects of the same sociological status. The sections of the city in which the subjects lived were very much the same, the congested districts around the large manufacturing centers. The writer has listed the occupations of the fathers of the boys just as they were given in the school records. No attempt has been made to classify the occupations. The reader may glance over the lists, and form his own opinion on the similarity or dissimilarity of the groups.

Seven of the fathers of the boys in the normal group were dead. The occupations of the fathers of the other boys were as follows,—4 potters, 3 wire-drawers, 3 laborers, 2 foremen over laborers, 2 bakers, 2 policemen, 2 machinists, 2 carpenters, 2 firemen, 2 tailors, 2 contractors, 2 merchants, and one each of the following, milkman, weight-master, decorator, janitor, superintendent of a pottery, manager of a chemical company, mason, hardware dealer, engraver, pattern-maker, shoe-maker, grocer, box-maker, laundry driver, printer, furnace tender, pipe fitter, railroad yard master, boilermaker, iron worker, brass worker, iron moulder and foundry worker.

Eight of the fathers of the boys in the retarded group were dead and two were invalids. The occupations of the fathers of the other boys were as follows, 7 laborers, 2 foremen over laborers, 8 potters, 2 peddlers, 2 contractors, 2 teamsters, 2 fish merchants, 2 blacksmiths, 2 painters, 2 masons, and one each of the following, tinsmith, watchman, fruit dealer, electrician, carpenter, plumber, carriage-maker, junk dealer, piano tuner, wire tinner, dairyman, meat packer, boilermaker, huckster, crockery dealer, railroader and machinist.

THE INFLUENCE OF THE LANGUAGE FACTOR.

The statement that the Binet tests depend on language training appears frequently in the literature of the subject. Investigators frequently refer the failure of their subjects on certain tests to deficiency in this sort of training or experience. Although investigators and critics of the scale frequently mention the language factor, no one has actually given a demonstration of the influence of this factor by comparing children of the same mental status but different linguistic training.

In this investigation it is necessary that this be done, for, in order to refer the differences found between the two groups to the relative diagnostic value of the tests, it must be shown that the differences are not due to the language training of the two groups. The subnormal group contains several children of non-English speaking parents. It is possible to compare this group with other subnormal children of the same age, who have been

in school the same length of time, and who are in the same grade, but who are children of English speaking parents. The two groups are objectively the same except that one group has had the advantage of the English language in the home, while the other has not. It is also possible to compare a group of normal boys of non-English speaking parents with normal boys of English speaking parents, the groups being similar in regard to age and school progress. If any differences occur between these groups, they may be referred to the influence of language training in the homes. The influence of the personal equation is absent, for the members of the groups were examined by one person at the same time. The influence of grade training is absent for the members of the groups compared have been in school the same number of years, and are in the same grades.

22 boys, aged 12, 13 and 14, of non-English speaking parents in the special classes were given the complete examination. They had been in school from 5 to 10 years, and were in grades II, III, IV, V and VI. To compare with this group, 22 boys of English speaking parents were selected who had been in school approximately the same length of time and were in the same grades. The average age of the 22 non-English retarded children was 13.26 yrs. ($MV=1.11$ yrs.) The average age of the 22 English retarded children was 13.10 yrs. ($MV=1.01$ yrs.) The average number of years in school of the non-English retarded children was 7.59 years. ($MV=1.05$ yrs.), of the English retarded group 7.09 yrs. ($MV=1.07$ yrs.) The average grade of both groups was the same, 3.68 ($MV=0.80$.) The average age of the non-English group is slightly higher, and they have been in school half a year longer on the average.

Twenty boys, aged 12, 13 and 14, of non-English speaking parents were examined in grades VI, VII and VIII. To compare with this group, 20 boys of the same age, grade and number of years in school, but of English speaking parents were selected. The average age of the non-English normal group was 13.58 yrs. ($MV=0.72$ yrs.). The average age of the English normal group was 13.53 yrs. ($MV=0.73$ yrs.). The average number of

years in school for both groups was the same, 7.30 yrs. ($MV=0.62$ yrs.). The average grade of both groups was the same, 7.45 ($MV=0.61$).

The results of the four groups of children were tabulated, and the percentage that each group passed each test calculated. The results are shown in Table 5. Column A gives the per cent. that each test was passed by the 40 children of English and non-English speaking parents combined. Column B gives the per cent. that the non-English group are above (+) or below (—) the English normal group. Column C gives the per cent. that each test was passed by the 44 retarded children of non-English and English speaking parents combined. Column D gives the per cent. that the non-English retarded group are above (+) or below (—) the English retarded group. Column E gives the difference between columns A and C or the per cent. that the 44 retarded pupils are above (+) or below (—) the 40 normal pupils.

TABLE 5.

Percentage Differences between Normal and Retarded Children of English and Non-English Speaking Parents.

	A	B	C	D	E
	% passed Normal.	Normal non-Eng. ±Eng.	% passed Retarded	Retarded non-Eng. ±Eng.	Retarded Normal.
Comparing remembered objects.....			97	— 6	
Counting backwards from 20 to 0.....			83	+22	
Indicating omissions in pictures.....			86	— 6	
Giving day and date.....	100	0	64	+16	—36
Enumerating the months.....	95	0	63	+ 8	—32
Naming the pieces of money.....	100	0	88	+13	—12
Making change	100	0	79	+22	—21
Arranging five weights.....	75	0	61	+ 5	—14
Copying designs from memory.....	65	—20	23	0	—42
Repeating five digits.....	100	0	95	0	— 5
Repeating seven digits.....	68	—25	27	— 9	—41
Using three words in a sentence. 2 ideas	100	0	59	+28	—41
Using three words in a sentence. 1 idea.	80	+20	43	+23	—37
Resisting suggestion	40	0	14	+27	—26
Naming 60 words in 3 minutes.....	90	+20	60	+33	—30
Giving 3 rhymes with "defender".....	0	0	0	0	0
Comprehending easy questions					
a. Train	100	0	98	— 5	— 2
b. Playmate	93	+ 5	61	— 6	—32
c. Broken	100	0	100	0	0
Any 2 out of 3.....	100	0	100	0	0

TABLE 5 (continued)

	A	B	C	D	E
Comprehending difficult questions					
a. Delayed	3	- 7	16	+ 4	+13
b. Important affair	88	+ 5	14	- 9	-74
c. Forgive easier	48	-15	9	- 9	-39
d. Asked opinion	98	+ 5	50	-10	-48
e. Actions vs. words	98	+ 5	20	-23	-78
Any 3 out of 5.....	88	+ 5	20	- 5	-68
Detecting absurdities in statements					
a. Bicycle rider	83	+ 5	36	-18	-47
b. 3 brothers	70	0	18	- 9	-52
c. Railroad accident	95	0	48	-32	-47
d. Suicide	93	- 5	64	0	-29
e. Friday unlucky	78	-15	11	-23	-67
Any 2 out of 3.....	93	- 5	30	-23	-63
Defining terms superior to use					
a. Fork	87	- 6	30	+ 6	-57
b. Table	80	-22	40	+16	-40
c. Chair	74	-12	37	+ 2	-37
d. Horse	85	- 1	26	+ 9	-59
e. Mother	67	-17	23	-21	-44
Any 3 out of 5.....	77	- 6	33	- 1	-44
Defining abstract terms					
a. Charity	45	+10	16	+ 4	-29
b. Justice	45	-20	5	0	-40
c. Kindness	48	-25	23	- 9	-25
Any 2 out of 3.....	48	-15	7	- 4	-41
Reconstructing dissected sentences					
a. early-hour	83	+15	9	0	-74
b. teacher-correct	95	0	39	+23	-56
c. dog-master	93	-15	32	+ 9	-61
Any 2 out of 3.....	95	-10	27	+18	-68
Repeating sentences of					
18 syllables	58	+ 5	31	-24	-27
20 syllables	15	-30	0	0	-15
22 syllables	3	- 5			
24 syllables	61	-37			
26 syllables	0	0	0	0	0
Solving problems from various facts					
a. Hanging from a limb.....	50	0	20	-23	-30
b. Neighbor's visitors	73	- 5	11	- 5	-62
Both correct	43	+ 5	7	-14	-36
Responses to pictures					
No. 1 { Description	70	-20	73	+18	+ 3
{ Interpretation	33	-15	30	+ 5	- 3
{ Emotion	3	- 5	0	0	- 3
No. 2 { Description	70	-20	66	+14	- 4
{ Interpretation	18	- 5	16	-14	- 2
{ Emotion	10	0	9	- 9	- 1
No. 3 { Description	70	-20	75	+22	+ 5
{ Interpretation	33	-15	30	+13	- 3
{ Emotion	3	+ 5	0	0	- 3
Summary of picture test					
Describing 2 out of 3 pictures					
(Age VII)	70	-20	75	+14	+ 5
Interpreting 2 out of 3 pictures					
(Age XV)	23	-15	20	- 5	- 3

The 62 differences between the normal English and the non-English groups, shown in column B, vary from -37% (i.e. 37% in favor of the English group) to $+20\%$ (i.e. 20% in favor of the non-English group). The median of the differences is 0% ($Q=7.5\%$). The average difference is -5.53% ($MV=9.33\%$). In the long run then the normal English group is about 5% above the normal non-English group.

The 63 differences between the retarded English and non-English groups, shown in column D, vary from -32% to $+33\%$. The median of the differences is 0% ($Q=11.5\%$). The average of the differences is $+0.83\%$ ($MV=11.22\%$). In the long run then the retarded non-English group are slightly better than the retarded English group.

From the figures given it is necessary to draw conclusions concerning what tests are influenced by language training, but this is very difficult on account of the lack of correspondence between the results of the two groups of English and non-English subjects. If one general cause, the language factor, were in operation in producing the divergencies in the results, then the results of the two groups should show a high correlation. The correlation (Spearman foot-rule method) of the differences between the normal English and non-English groups and the retarded English and the non-English groups (i.e. the correlation between columns B and D) is -0.06 ($pe=0.055$) or no correlation. This would indicate that the differences were due to chance rather than to the one general factor, language training.

Inasmuch as the number of subjects (20 or 22) in each group is small, there is a very strong possibility that the differences might be due to chance. A glance at columns B and D shows that in some cases the results agree, and in other cases they are exactly opposite. In the test of defining terms superior to use, for example, the non-English group are 17% below the English group in the case of the normal subjects and 21% in the case of retarded subjects in defining "Mother." In defining "Table," however, they are 22% below in the normal group but 16% above in the retarded group. The frequency of occurrence of

like and unlike signs in columns B and D is as follows: oo occurs 5 times; o+, 10 times; o—, 10 times; +—, 21 times; ++, 3 times and ——, 11 times. If the results were due to chance, the number of unlike signs should be twice that of like signs. The actual number of like signs (oo, ++, and ——) is 19, and the actual number of unlike signs (o—, o+ and +—) is 41, the proportion expected. Taking only the plus and minus signs, they are like in 14 cases and unlike in 21 cases. These figures would also indicate that the results were due to chance, but they are not conclusive, for the magnitudes of the differences are not taken into consideration.

The results in general would indicate that no one factor was in operation in producing the differences found, but the possibility remains that the language factor might influence some of the individual tests. If this factor were in operation on the individual tests, then the sum of the differences between the normal English and non-English groups and the retarded English and non-English groups would indicate what tests were influenced, for those differences with unlike signs would cancel out in combination, while the differences with like signs or a common tendency would be exaggerated. Combining the differences in this manner (i.e. taking the algebraic sum of the differences shown in columns B and D) the 60 sums of differences obtained vary in magnitude from —38% to +53%. The average of the sums of differences is —4.32% (MV = 13.29%). There is no method of deciding which of the tests are influenced by language training. The lists of tests are given below so that the reader may form his own opinion. The tests found in the highest 25% in favor of the non-English groups, and in the highest 25% in favor of the English groups are given.

THE 15 TESTS SHOWING THE LARGEST SUMS OF DIFFERENCES
IN FAVOR OF THE ENGLISH GROUPS.

- 38% Absurdity e. (Friday unlucky)
- 38% Definition e. (Mother)
- 34% Abstract definition c. (Kindness)
- 34% Repeating 7 digits.
- 32% Absurdity c. (Railroad accident)

- 30% Repeating sentence of 20 syllables.
- 28% Passing 3 out of 5 absurdities.
- 24% Comprehension c (forgive easier).
- 23% Problem a (Hanging from a limb).
- 20% Interpreting 2 out of 3 pictures.
- 20% Abstract definition b (Justice).
- 20% Copying designs from memory.
- 19% Interpreting picture 2.
- 19% Repeating a sentence of 18 syllables.
- 19% Passing 2 out of 3 abstract definitions.

THE 15 TESTS SHOWING THE LARGEST SUMS OF DIFFERENCES
IN FAVOR OF THE NON-ENGLISH GROUPS.

- +53% Giving 60 words.
- +43% 3 words in sentence (1 idea).
- +28% 3 words in a sentence (2 ideas).
- +27% Resisting suggestion.
- +23% Dissected sentence b (teacher-correct).
- +22% Making change.
- +16% Naming date.
- +15% Dissected sentence a (early-hour).
- +14% Abstract definition a (Charity).
- +13% Naming money.
- + 8% Definition d (horse).
- + 8% Passing 2 out of 3 dissected sentences.
- + 8% Enumerating the months.
- + 5% Emotional interpretation of picture 3.
- + 5% Arranging 5 weights.

The problem now arises of how large a difference may be taken to indicate the influence of the language factor. This problem has no answer outside of personal opinion. Taking the differences in favor of the English groups, the test of copying the designs from memory shows a difference of 20% in favor of these groups, but it is hard to see how this test can be influenced by language training. Again, the test of repeating 7 digits shows a difference of 34% in favor of the English groups. It is also hard to understand why this test should involve the language factor as both groups had on an average over seven years of experience in using digits. Taking the results that are in favor of the non-English speaking groups, the date test shows these children ahead 16%, the test of making change 22% and

the line suggestion test 27%. It is difficult to understand how any of these tests may be influenced by language training. Using the tests enumerated as a limit, it might be concluded that some of the absurdities tests and some of the definitions tests show the influence of language training in favoring children of English speaking parents. If this conclusion is drawn, however, it is also necessary to conclude that the tests of naming 60 words and constructing a sentence from three given words show a larger influence of language training in favoring the children of non-English speaking parents. This conclusion is certainly possible, for the training of this group of subjects in two languages may be a positive help. The reader may draw his own conclusions.'

Although no definite conclusions may be drawn concerning the presence of the language factor, it is possible to estimate the importance of this factor in the present investigation by comparing it with another factor, that of the intellectual differences between the groups. Columns A and C show the per cent. that the 40 normal children and the 44 retarded children (English and non-English combined) pass the individual tests, and column E indicates the percentage difference between these groups. The 60 differences between the normal and retarded group vary from -78% (i. e. 78% in favor of the normals) to +13%. The average difference is -30.48% ($MV=20.72\%$). The largest difference between the English and non-English groups is 37% or only 7% higher than the average difference between the normal and retarded groups. The average difference between the English and non-English normal groups is about 5% or one sixth of the average difference between normal and retarded children. It is therefore possible to conclude that the language factor has very little importance as compared to the intellectual differences between the groups.

In the present investigation, the conclusion that the language factor has very little importance as compared to the factor of the intellectual differences between the two groups does not mean that the language factor plays no part in the Binet tests, but that it may be disregarded in this study. It was not possible to demonstrate whether the differences found between the language groups were due to chance or training. The absence of

correlation between the results of the retarded and normal language groups would indicate that the differences were due to chance. Certain of the individual tests may show the influence of this factor, but this influence may be disregarded for it should be the same in both groups, the normal and retarded. Any effect of language training would be equal in both groups, for both groups contain approximately the same number of children of non-English speaking parents, both groups come from the same sort of homes, and both groups have had the same amount of linguistic training in the homes and in the schools. The differences may be related to the intellectual differences between the groups, and these differences may therefore be used as indices of the diagnostic value of the tests.

V. DIAGNOSTIC VALUE OF BINET TESTS

In the Introduction it was shown that the Binet scale is not a reliable instrument for diagnosing the higher grades of mental defect, for, as suggested by Binet (7) and demonstrated by Descoedres (20) and Chotzen (18), it is composed of some tests that are effective and others that are ineffective in diagnosing intelligence. To determine what tests were most effective, the writer gave the tests to two groups of children who had had the same physical and environmental opportunities but who showed different school progress. It was not possible to account for the fact that one group had progressed only half as far in school as the other without assuming a difference in the general intellectual endowment of the groups. The differences in the reactions of the groups to the tests could not be due to differences in sociological status or school training, nor could the personal equation or the language factor influence the results to any extent. These differences may therefore be referred to the intellectual differences between the groups, and the magnitude of the differences used as indices of the diagnostic value or effectiveness of the tests.

The logic of this method of measuring the effectiveness of the individual tests is the same as that of the method which Binet (5) proposed for estimating the importance of stigmata in the diagnosis of subnormality. According to Binet's proposed method of calculation, if a certain stigma were always found among subnormals and never among normal individuals, this stigma would have the value of 100%. Another stigma found among all subnormals and 50% of normals would have a value of 50%. In this way the principle of calculation which Binet proposed would attach to each stigma its "coefficient d'importance," and the relative certainty of these diagnostic indices would be measured on a scale of 100. The present investigation simply reverses the process. If a certain test ability were present in all normals and not present in any subnormals, its diagnostic value would be

100%. If any ability occurs with the same frequency in normals and subnormals, the percentage performance of each group would be the same and its diagnostic value would be 0.

The per cent. that each test was passed by the 58 normal and 59 retarded subjects, and the differences between these percentages, or the diagnostic value of each test, are shown in table 6.

TABLE 6.
Diagnostic Value of Each Test.

	Per cent. passed by normal	Per cent. passed by retarded	Diagnostic value
Giving day and date.....	100	71	-29
Enumerating the months.....	98	55	-43
Naming the pieces of money.....	100	91	-9
Making change	98	83	-15
Arranging five weights	74	56	-18
Copying designs from memory.....	67	31	-36
Repeating five digits	100	98	-2
Repeating seven digits	64	34	-30
Using three words in a sentence (2 ideas).....	100	71	-29
Using three words in a sentence (1 idea).....	79	56	-23
Resisting suggestion	33	21	-12
Naming 60 words in 3 minutes.....	93	66	-27
Giving three rhymes with defender.....	0	0	0
Repeating a sentence of 26 syllables.....	0	2	+2
Comprehending easy questions			
a. Train	100	98	-2
b. Playmate	95	72	-23
c. Broken	100	100	0
*Any 2 out of 3.....	100	100	0
Comprehending difficult questions			
a. Delayed	7	17	+10
b. Important affair	88	12	-76
c. Forgive easier	50	14	-36
d. Asked opinion	98	47	-51
e. Actions vs. words	91	17	-74
*Any 3 out of 5.....	86	15	-71
Detecting absurdities in statements			
a. Bicycle rider	90	48	-42
b. 3 brothers	71	25	-46
c. Railroad accident	95	54	-41
d. Suicide	91	68	-23
e. Friday unlucky	81	19	-62
*Any 3 out of 5.....	95	42	-53
Defining terms superior to use			
a. Fork	84	35	-49
b. Table	84	46	-38
c. Chair	81	39	-42
d. Horse	86	33	-53
e. Mother	72	29	-43
*Any 3 out of 5.....	84	33	-51

TABLE 6 (continued)

	Per cent. passed by normal	Per cent. passed by retarded	Diagnostic value
Defining abstract terms			
a. Charity	53	12	-41
b. Justice	47	5	-42
c. Kindness	48	24	-24
*Any 2 out of 3.....	59	8	-51
Reconstructing dissected sentences			
a. early-hour	84	14	-70
b. teacher-correct	93	39	-54
c. dog-master	97	32	-65
*Any 2 out of 3.....	100	29	-71
Solving problems from various facts			
a. Hanging from a limb.....	53	32	-21
b. Neighbor's visitors	74	23	-51
Both correct	45	10	-35
Responses to pictures			
No. 1 { Description	74	71	- 3
{ Interpretation	33	24	- 9
{ Emotion	2	2	0
No. 2 { Description	69	73	+ 4
{ Interpretation	19	12	- 7
{ Emotion	12	10	- 2
No. 3 { Description	74	78	+ 4
{ Interpretation	28	34	+ 6
{ Emotion	3	0	- 3
Summary of pictures test			
Describing 2 out of 3 pictures (Age VII)...	74	76	+ 2
Interpreting 2 out of 3 pictures (Age XV)...	21	15	- 6

A glance at table 6 shows the large variation in the diagnostic value of the tests. The tests of reconstructing dissected sentences, for example, show differences between the groups of 54%, 65%, 70% and 71%, while the test of resisting suggestion shows a difference of but 12%. The first test is passed by all the normal group, but by only 29% of the retarded group. The second test is passed by 33% of the normal and 21% of the retarded groups. Both of these tests are "twelve year" tests. The suggestion test, although quantitatively equal to the sentence test, is almost equally as difficult for both groups, while the sentence test is universally passed by one group and is difficult for the other. These figures would show then that the test of recon-

*The scores "Any 2 out of 3" or "Any 3 out of 5" refer to Binet's method of counting certain tests passed if the subject passes a certain number of the parts of the test.

structing dissected sentences is highly diagnostic of intelligence, while the line suggestion test has no value as an intelligence test.

The differences between the normal and retarded groups vary from -76% to $+10\%$, the median being -29% ($Q=23\%$). At first thought it might seem that the tests could be arranged immediately in the order of their diagnostic value on the basis of the figures given in table 6. More careful study shows that the method, although conclusive concerning certain tests, is inconclusive concerning others. The truth of this statement is very clearly shown by referring to the tests that stand out at the extremes of the list—the tests that show the highest and the lowest diagnostic values.

The following list contains all the tests that show a diagnostic value over 50% :

- 76 Comprehension b (Important affair).
- 74 Comprehension e (Actions vs. words).
- 71 Any 3 out of 5 comprehension questions.
- 71 Any 2 out of 3 dissected sentences.
- 70 Dissected sentence a (early-hour).
- 65 Dissected sentence c (dog-master).
- 62 Absurdity e (Friday unlucky).
- 54 Dissected sentence b (teacher-correct).
- 53 Any 3 out of 5 absurdities.
- 53 Definition d (Horse).
- 51 Any 3 out of 5 definitions superior to use.
- 51 Problem b (Neighbor's visitors).
- 51 Any 2 out of 3 definitions of abstract terms.
- 51 Comprehension d (Asked opinion).

The 14 tests in the above list are parts of six tests, comprehending difficult questions, reconstructing dissected sentences, detecting absurdities, defining in terms superior to use, defining abstract terms and solving problems. Turning to these tests in table 6, it is seen that their diagnostic value is in general very high. This would indicate then that these tests were the most effective ones in the scale for differentiating the groups in question.

Conclusions as definite can not be drawn concerning all the tests at the other extreme. The following list contains the tests that show a diagnostic value under 10% :

- + 10 Comprehension a (Delayed).
- + 6*Interpretation. Picture 3.
- + 4*Description. Picture 3.
- + 4*Description. Picture 2.
- + 2*Describing 2 out of 3 pictures (Age VII).
- + 2 Repeating a sentence of 26 syllables.
 - o Giving 3 rhymes with defender.
 - o Easy comprehension c (Broken).
 - o Any 2 out of 3 easy comprehension.
- o*Emotion. Picture 1.
- 2*Emotion. Picture 2.
- 2 Easy comprehension a (Train).
- 2 Repeating 5 digits.
- 3*Description. Picture 1.
- 3*Emotion. Picture 3.
- 6*Interpreting 2 out of 3 pictures.
- 7*Interpretation. Picture 2.
- 9*Interpretation. Picture 1.
- 9 Naming the pieces of money.

The above list of 19 tests contains 11 tests of one sort, so that there is strong evidence that this test, describing and interpreting pictures, has no value in diagnosing intelligence. It is not possible to draw conclusions concerning all the other tests in the list, because it is not possible to determine the relation between the difficulty of a test and its diagnostic value. If, for example, the members of the groups studied had been asked if they were little boys or little girls, 100% of both groups would have passed, and the diagnostic value of the test would have been zero, or, if they had been asked to translate a passage of Greek, none of them would have passed, and the diagnostic value would be zero again. The tests of naming the pieces of money and comprehending easy questions show no diagnostic value, but that does not prove that these tests would have no diagnostic value in differentiating groups with less ability than the ones examined. In the same way, the tests of repeating a long sentence and giving rhymes with "defender" are like the passage in Greek, and the fact that they show no diagnostic value for these groups does not prove that they would not be effective in differentiating groups of higher intelligence than the ones examined.

Between these extremes of tests that are entirely below or entirely above the range of ability of the groups examined the tests may be distributed in more or less uniform steps. Some tests may be just above the lowest ability of the groups, others just below the highest ability, and others nearer the median of these extremes. Concerning every test it is possible to raise the question of how the diagnostic value would be changed if the test had been more or less difficult, or how this value would be changed if the groups to whom it had been given had been more or less intelligent. It is not possible to alter the difficulty of the tests after they have been given, but in two cases it is possible to change the passing mark, and to calculate the per cent. that would have passed had the passing mark been more or less severe.

In the 60 word test, a subject is required to give at least 60 words in three minutes to pass the test. As a matter of fact, the subjects gave anywhere from 31 to 196 words in the required time. The retarded group gave from 31 to 148 words, the median being 69 ($Q=15.5$). The normal group gave from 43 to 196 words, the median being 83 ($Q=11.5$). If the passing mark had been 75 words instead of 60 words, 77% of the normals and 37% of the retarded would have passed, and the diagnostic value would have been 40% instead of 27% as shown in table 6. In this way it is possible to calculate the percentage passed and the diagnostic value for each passing mark. These values are shown for 12 passing marks in table 7. The test of copying designs from memory was scored according to the arbitrary point system described in chapter III of the preceding section. The scores of both groups varied from 0 to 20 points, the median of the retarded being 10 ($Q=6$), and of the normal 18 ($Q=4.5$). The percentage passed and the diagnostic value for 10 passing marks are shown in table 7.

The diagnostic value of the 60 word test rises from 10% (where the passing mark is 110 words and the test too hard) to 40% and falls again to 12% and eventually zero when the test is too easy. The diagnostic value of the design test rises from 23% when the test is too difficult to 36% and down to zero when the test is too easy. Every test then has a value which will be

TABLE 7.

Relation between the Difficulty of Two Tests and their Diagnostic Value.

Naming 60 words in 3 minutes				Copying designs from memory			
Passing mark.	Per cent. passed by normal	Per cent. passed by retarded	Diagnostic value	Passing mark.	Per cent. passed by normal	Per cent. passed by retarded	Diagnostic value
No. of words	group	group		No. of points	group	group	
110	14	4	-10	20	26	3	-23
100	19	7	-12	19	34	9	-25
95	31	14	-17	17	55	21	-34
90	38	16	-22	15	67	31	-36
85	47	19	-28	13	71	41	-30
80	64	30	-34	10	76	52	-24
75	77	37	-40	7	83	60	-23
70	84	47	-37	5	88	71	-17
65	88	53	-35	3	98	88	-10
60	93	66	-27	0	100	100	0
50	97	77	-20				
40	100	88	-12				

called the *Maximum Diagnostic Value*. This value is 40% for the 60 word test, and 36% for the designs test.

In the two tests discussed, the groups were constant and the difficulty of the tests varied. In cases where the difficulty of the test may be varied, the Maximum Diagnostic Value may be obtained, provided of course that the method of scoring the test is an accurate expression of the intellectual factors involved. In other tests it is not possible to alter the difficulty, and the Maximum Diagnostic Value can not be determined unless the test be given to groups of varying degrees of intelligence. In this case, the test is constant and the intellectual level of the groups must be varied in order to find the Maximum Diagnostic Value. In this experiment, the groups are constant, so that the method, although conclusive in regard to certain tests, is inconclusive in regard to others. It is conclusive in regard to tests that show a high diagnostic value, and those in which the scoring allows of the determination of the Maximum Diagnostic Value, but inconclusive concerning most of the tests that show no diagnostic value. The method admits of many positive, but few negative conclusions.

It is possible to use two measures of the diagnostic value of the tests, the absolute difference between the per cent. that the normal

and retarded groups pass each test, and the relative difference between these percentages. The dissected sentence test was passed by 100% of the normal group and 29% of the retarded group, the five weights test by 74% normal and 56% retarded, the test of defining "charity" by 53% normal and 12% retarded, and the test of giving an intellectual interpretation of picture 1 by 33% normal and 24% retarded, making the absolute differences between the groups 71%, 18%, 41% and 9% respectively. The difficulty of the tests for the normal group varied as shown by the percentages passed, 100%, 74%, 53% and 33%. Had all of the tests been equally within the range of the groups, it is possible that the diagnostic values would have been different. The relative measure would be the per cent. that the absolute difference was of the per cent. passed by normals, the values in the case of the four tests cited being 71%, 24%, 77% and 27%.

The use of the relative differences would imply that the diagnostic values would have changed if the intellectual level of the groups had been lower or higher, just as the diagnostic value varies if the difficulty of the test (the passing mark) is raised or lowered. This would undoubtedly have been the case, yet we are not warranted in making inferences from the performance of the two groups tested to the performance of any other groups. The data from the two groups give no information concerning the growth of abilities with age or with intelligence. The percentages merely indicate the actual performance of the groups tested. The absolute differences are used as measures of the diagnostic values of the tests in this study, and contain no implications concerning the nature of the performance of other groups.

The diagnostic value of the tests is not the only criterion that the figures in table 6 afford for judging the relative merits of the individual tests. The tests used were in the VII, VIII, IX, X, XII and XV year groups. Inasmuch as the normal subjects were 12, 13 and 14 years of age, it is legitimate to expect that all or practically all of these subjects should pass the VII, VIII, IX and X year tests. The VII year test of describing pictures is failed by 26% of the normal group so that something would ap-

pear to be wrong with this test. The VIII and IX year tests are almost universally passed with the exception of the definitions test which is failed by 16% of the normal group. In the X year group, the absurdity and sentence tests are almost universally passed, the comprehension test is failed by 14%, the 5 weights test by 26%, and the design test by 33%.

The normal group is composed of 18 subjects aged 12 and 20 subjects of 13 and 14. According to Binet's procedure in calibrating the tests for the different years, about 70 or 75% of the 12 year and all of the 13 and 14 year normal boys should pass the XII year tests, or approximately 90% of all the group should pass. This percentage is approximated by the sentence test (79%), the 60 word test (93%) and the dissected sentence test (100%). The definitions test would appear to be too difficult (59%) and the suggestion test entirely too difficult (33%).

If the theoretical curve of growth of the XV year tests could be guessed at, it would probably approximate 0% at 12, 25% at 13, 50% at 14, 75% at 15 and 100% at 16. Roughly then it would be expected that about 25% of the 12, 13 and 14 year normal children would pass these tests. None of these subjects passes the tests of repeating the long sentence or giving rhymes, 21% interpret pictures, 45% solve the problems and 64% repeat 7 digits. The results of the normal group should probably not be used to criticize the "fifteen year" tests, for these subjects were all under 15. The writer has available the results of 10 boys in the Princeton High School, three of whom were 15, one 16, four 17 and two 18 years of age. Three of these boys failed to repeat 7 digits and to interpret pictures, four failed to repeat the long sentence and only one gave three rhymes with "defender." The writer has given these tests to many normal adults but has not recorded them systematically. He has available however the results of seven graduate students all of whom had at least a bachelor's degree. All 7 repeated the 7 digits and the sentence of 26 syllables, two failed to give 3 rhymes, and one gave the "three year" response of enumerating the objects in the pictures, failing in VII and XV. These groups of subjects also throw interesting side lights on other tests. Of the 10 high school

students, 2 failed to copy the designs, 2 failed to arrange five weights, and 6 failed the line suggestion test. Of the 7 advanced university students, 1 failed the design test, 1 failed the weights test and 2 failed the line suggestion test. If these tests are for "ten" and "twelve" year mentality it is right to expect that all of these groups of subjects should pass them.

The figures given in table 6 also show whether the different sub-questions under the various tests are of the same difficulty. In general the results of the sub-questions are about the same. The most marked exceptions appear in the test of comprehending difficult questions, where question *a* is practically impossible, and questions *b*, *d* and *e* almost twice as easy as question *c*.

For the convenience of the reader, the tests shown in table 6 are arranged in table 8 in the order of their diagnostic value as shown by this method of study.

TABLE 8.

Per Cent that Normal Group Pass Each Test.

(Tests Arranged in the Order of Their Diagnostic Value.)

	Diagnostic value	Per cent. passed by normal
1. Comprehending difficult questions (3 out of 5).....	-71	86
2. Reconstructing dissected sentences (2 out of 3).....	-71	100
3. Detecting absurdities in statements (3 out of 5).....	-53	95
4. Defining in terms superior to use (3 out of 5).....	-51	84
5. Defining abstract terms (2 out of 3).....	-51	59
6. Enumerating the months.....	-43	98
7. Copying designs from memory.....	-36	67
8. Solving (both) problems from various facts.....	-35	45
9. Repeating 7 digits.....	-30	64
10. Giving the day and date.....	-29	100
11. Using 3 words in a sentence (2 ideas).....	-29	100
12. Naming 60 words in 3 minutes.....	-27	93
13. Using 3 words in a sentence (1 idea).....	-23	79
14. Arranging 5 weights.....	-18	74
15. Making change.....	-15	98
16. Resisting suggestion.....	-12	33
17. Naming the pieces of money.....	-9	100
18. Interpreting 2 out of 3 pictures (Age XV).....	-6	21
19. Repeating 5 digits.....	-2	100
20. Giving 3 rhymes with "defender".....	0	0
21. Comprehending easy questions.....	0	100
22. Repeating a sentence of 26 syllables.....	+2	0
23. Describing 2 out of 3 pictures (Age VII).....	+2	74

Describing and Interpreting Pictures. These tests appear near the bottom of the list. The small diagnostic value of these tests is surprising in view of the fact that Binet considered the test of interpreting pictures one of the most important in differentiating normals from morons. Binet considered this the most valuable test in the scale. "We place it above all others; and were we limited to one test we would without hesitation choose this one." (Town's (72) translation, page 13.) The results of this investigation would show that there is something radically wrong with this test. Very few of the subjects pass it (21% of the normal group) but this is to be expected if it is a "Fifteen year" test. Three of the high school students failed it, and one of the college adults gave the characteristic "three year" response. Furthermore the results show that the retarded children are just as likely to give an emotional or intellectual interpretation as the normal children. The same holds of the test of describing pictures which shows no difference between the groups. In fact 26% of the normal group fail to describe the pictures and give the enumeration response which is characteristic of the "three year" level. The pictures are not of the same difficulty (see table 6), it being easier to give an intellectual interpretation of pictures 1 and 3 than of picture 2, while the emotional interpretation of the latter appears more frequently.

The explanation of the fact that this test shows no diagnostic value probably lies in the instructions. According to Binet's procedure the child is given the picture and asked "What is this?" If he says "It is a picture," the question is put "Tell me what you see there." In this experiment, the instructions were "Tell me what you see in this picture." The use of the word "what" probably induces the response by enumeration. At any rate the instructions do not seem to produce the same "Aufgabe" in all subjects, and it is most important that all subjects should have the same "Aufgabe" on every test. Persons able to interpret the pictures do not interpret them because they think something else is expected of them. In most of the cases the subjects' responses are not real measures of their ability. The writer believes that the ability to give an intellectual or emotional interpretation of pictures is diagnostic of intelligence. If this factor is correlated with intelligence, it would seem reasonable to try to test for it. This could be done fairly accurately by using more than three pictures, and by framing the questions so as to demand the answer by interpretation, saying "What has happened here?" or "What is the matter with these people?" The test as it stands now is worthless.

Repeating a Sentence of 26 Syllables. This test shows no diagnostic value because it is too difficult. None of the normal subjects and only one retarded subject passed this test. Four of the high school students failed it, so that it would appear too difficult for a "fifteen year" test. In this test a graded series of 12 sentences (11 of them from Town and 1 from Whipple's (75) manual) were given, the sentences varying in length from 10 to 32 syllables. The procedure used was that of starting within the subject's range and continuing up the scale until he had failed two sentences in succession, the number of syllables in the last sentence being taken as the measure of his ability in the test. When working with normal subjects, this

procedure was found to be inadequate for some subjects would fail two or three sentences in succession and then pass the next.

Taking the results of the normal subjects, in one case a subject failed four sentences in succession and then passed the next. In 19% of the cases the subjects failed three sentences in succession and passed the next, and in 45% of the cases they failed two in succession and passed the next. The procedure was entirely wrong then in taking the subject's threshold as the point beyond which he failed two tests. The error made by the writer was that of considering the order of the increasing number of syllables to be the measure of the increasing difficulty of the test. This was not the case. The tests of repeating sentences of 18, 20, 22 and 24 syllables were given to the normal group 90% of the possible number of times. 51% passed 18 syllables, 18% passed 20 syllables, 3% passed 22 syllables and 61% passed 24 syllables. The order of difficulty is therefore 24, 18, 20, 22. Of the 61% of the subjects who passed the 24 syllable sentence, 6% failed to repeat 16 syllables, 41% failed to repeat 18 syllables, 79% failed 20 syllables and 94% failed 22 syllables. The explanation of this probably lies in the fact that the 24 syllable sentence was logically simpler than the others. The factor of logical memory can be separated from tests of auditory memory only by the use of nonsense syllables. The 24 syllable sentence was not the only one in error, however. One subject repeated 22 syllables and failed to repeat 16, 18 and 20 syllables, while 17% of those who repeated 20 syllables failed to repeat 18 syllables. The order of increasing number of syllables is therefore not the order of increasing difficulty.

According to the procedure, the test was discontinued as soon as two sentences in succession had been failed. On this account, all of the sentences were not given all of the possible number of times, and, as the error in procedure was not discovered till normal subjects were examined, a different range of sentences were given to the two groups, so that it is not possible to compare their results on many sentences. The 26 syllable sentence proved too difficult for both groups. The 24 syllable sentence was only given to one member of the retarded group so that no comparison is possible. The 22 syllable sentence was given to all the normal subjects and only passed once, while it was passed but once by the 17 retarded subjects to whom it was given, and is therefore too difficult to show any diagnostic value.

Some indication of the diagnostic value of the 16, 18 and 20 syllable sentences may be obtained. The 20 syllable sentence was given to 75% of the retarded and 98% of the normals. It was passed by 2% of the retarded and 18% of the normals, making the diagnostic value 16%. The 18 syllable sentence was given to 90% of the retarded and 91% of the normals. It was passed by 28% of the former and 51% of the latter, making the diagnostic value 23%. The 16 syllable test was given to 81% of the retarded and 47% of the normals, it being assumed that all the subjects to whom the test was not given would have passed if the test had been given. The test was actually passed by 67% of the retarded and 74% of the normals, but if all the subjects had passed whom it was assumed would pass, 72% of the retarded and 88% of the normals would have passed. The diagnostic value of this sentence is therefore between 7% and 16%. The 14 syllable sentence was

too easy to show any diagnostic value. The number of syllables for the 16, 18, 20 and 22 syllable sentences is the correct measure of their difficulty as shown by the per cent. that normals passed them (74%, 51%, 18% and 3%). The 14 and 16 syllable sentences are too far below the ability of the groups to show any diagnostic value, and the 20 and 22 syllable sentences too far above this ability. The diagnostic value shown by the 18 syllable sentence (23%) may therefore be taken as the Maximum Diagnostic Value for the test of repeating sentences.

Comprehending Easy Questions. Nothing can be said concerning the diagnostic value of this test, because all members of both groups passed two of the three questions. Question b (Playmate) is more difficult than the other two.

Giving Three Rhymes with "Defender." This test shows no diagnostic value, because none of the normal or retarded subjects passed the test. The writer's experience has been that this test is practically impossible for any but exceptionally gifted adults, and is not a fair test of "fifteen year" mentality. The error lies in considering the process of finding three tri-syllabic English words ending in "ender" equal in difficulty to the process of finding three French words ending in "ance" (the Binet test word being "obéissance"). It would be better to admit that the test can not be translated. The normal subjects gave in all 30 rhymes with "defender" and the retarded subjects 16. If the subjects who succeeded in giving one or more words are considered as passing the test, 7 of the retarded and 26 of the normals passed, making the diagnostic value 33%.

Before asking the subjects to give rhymes with "defender," a practice test was given in which the words "day," "mill" and "spring" were used, half a minute being allowed for each word. The total number of rhymes given by both groups for "day" was 439, for "mill" 457 and for "spring" 273. The latter word is therefore much more difficult. The difference in the number of rhymes given by the two groups for "day" was 73, for "mill" 111 and for "spring" 65. The word "mill" would seem to have the highest diagnostic value. Taking the total number of rhymes given for all three words as the measure of ability, the normal subjects varied from 0 to 23, the median being 12.5 ($Q = 3.75$). The retarded group varied from 0 to 21, the median being 9 ($Q = 5.25$). Calculating the percentages of each group that would have passed had the passing mark been fixed at any number of words from 0 to 23, and subtracting to determine the diagnostic value at each passing mark, the Maximum Diagnostic Value for this test was 32% at the passing mark of 5 or 6 words. The value is very close to the value found for giving one or more rhymes with "defender" (33%), so that these figures probably express the general value of rhyming tests in differentiating the intellectual differences between the groups. Binet included the rhyming test in the list of six tests that he considered valuable in differentiating morons from normals.

Repeating Five Digits. This test shows no diagnostic value, because it is too far within the ability of the groups.

Repeating Seven Digits. This test shows a diagnostic value of 30%.

Naming the Pieces of Money. No conclusions may be drawn concerning

the differential value of this test, as it was failed by only 5% of the retarded group.

Making Change. Although this test is slightly more difficult for the retarded group than naming money, it is still too far within the ability of the groups to show its true diagnostic value.

Resisting Suggestion. This test shows practically no diagnostic value (12%), and is passed by only 33% of the normal group. This test admits of a more accurate scoring, inasmuch as there are three lines on which judgments must be made. The normal group in all gave 58 correct judgments out of 174 possible judgments or 33% correct. The retarded group gave 45 correct judgments out of 174 or 26%, making the diagnostic value 7%. The small percentage passed by the normal group is surprising in view of the fact that this test is a "twelve year" test. 6 of the 10 high school students failed this test and 2 of the adults. It is certainly not a test for "twelve years" then, and the writer doubts if it is a test for intelligence, as in his experience persons of ability fail it just as readily as persons of no ability. It is seen to be equally difficult for retarded and normal subjects. Schmitt (57) notes two types of failure on this test, the typical type of failure according to Binet of accepting the suggestion of the first three lines, and the failure due to the fact that the subject actually judges the lines unequal after studying them. In the writer's opinion, this analysis is correct. It means that intelligent persons may fail the test by actually misjudging the length of the lines—cases in which the factor of suggestion is entirely absent. Even if suggestion does influence intelligent subjects in this test, it is not a symptom of defective intelligence to have suggestion warp one's judgment on sensory data. An experiment on suggestion by means of the size weight illusion conducted by Dresslar (24) actually showed that the brighter children were more suggestible than the duller children. Terman (65) has eliminated this test from the latest Stanford revision. Persons who have conducted laboratory experiments on the thresholds of sensation know how difficult it is to rule suggestion out even with highly trained subjects. In doubtful cases of discriminative judgments the subject is apt to take any clue. As a general rule the intelligent person is quite ready to discredit the evidence of his senses. Very rarely will he discredit the conclusion of a reasoning process, however, and the influence of suggestion is a symptom of defect only when it warps one's intellectual judgments.

Arranging Five Weights. This test shows a low diagnostic value (18%). A more accurate method of scoring (taking account of the actual number of successes and failures) shows the normal group arranging the weights correctly in 70% of their 174 trials, and the retarded group in 55% of their 176 trials, making the diagnostic value 15%. This result is surprising in view of the fact that Binet included this test in the list of six tests that he considered most valuable for differentiating morons from normal individuals. Binet styles it "An excellent test which presupposes no schooling or acquired knowledge, and expresses intelligence in its most natural form," and says that "this test is one of those which best detect intelligence without culture, as it is absolutely independent of all instruction." (Town's translation,

pages 41 and 42.) The results of this investigation would show that the test as it stands is also independent of intelligence.

The reason for this lack of correlation with intelligence lies in the indiscriminate mixing of the many factors involved in the test. In discussing this test, Binet points out the various types of response—"Many children do not understand the explanation and remain motionless; so much the worse for them. Others place the boxes in any order without lifting them; and from the little attention that they give them, it is easy to see that they make no comparison. Others understand that the heaviest must be placed first; and they distinguish between the weights of the others most accurately, but they are incapable of arranging the other boxes in the order of their decreasing weight; this idea of decreasing weight is unintelligible to them. They do not lack in sensibility to weight, but in ability to arrange. Others finally grasp the idea of decreasing order, and they come a little nearer to applying it; they arrange such series as 15, 12, 9, 3, 6, where a single box is misplaced; they can do better, they fail from lack of attention and care. This is not a grave error. Nevertheless, we exact two absolutely correct arrangements." (Town's translation, page 42.) Other writers are in general agreement with Binet's analysis of the factors involved. Yerkes (82) classifies the factors as "Kinaesthetic discrimination, ideation (notion of series), attention."

In this experiment a control test of five definitely supraliminal weights (20, 30, 45, 70 and 100 gms.) was used. This test was passed by all the normal and retarded subjects showing that the intellectual ability to comprehend a serial arrangement or to make the logically necessary comparisons was present. The failures were therefore failures in sensory discrimination, and were of the sort that Binet characterized as "not grave." The results show then that the test as a test of sensory discrimination has no diagnostic value. Another proof of the test's lack of worth is the fact that 26% of normal 12, 13 and 14 year boys fail to pass what is supposedly a "ten year" test. Schmitt reports that half of a college class of twenty students failed to arrange the five weights correctly, and believes that "the grasp of the idea of arranging them serially, and an intelligent attempt to do so, is the significant part of the test." (Page 39.) In the writer's experience, the control test of 20, 30, 45, 70 and 100 gram weights proved to be very useful and highly diagnostic of the deficiency in the intelligence of low grade cases. It would seem then that the intellectual factors of comprehending a serial arrangement and making the logically necessary comparisons were diagnostic of intelligence while the sensory discrimination was not.

The above conclusion differs little from those of Peterson and Doll (51) who find that the sensory capacity of defective children in muscle sense is not noticeably below normal, and that the slight differences found may be accounted for on an intellectual rather than a sensory basis. They affirm that the test of discriminating lifted weights which they used was not diagnostically valuable except in types of success in following instructions. Smith (59) reports a high correlation between pitch discrimination and general intelligence, but believes that this correlation is due to the intellectual factors in the test rather than to any physiological factors of sensory dis-

crimination. These two factors, the intellectual and the physiological (or as Smith calls them, the "elemental"), enter into almost all sensory tests, and the writer believes that most of the correlations reported between sensory discrimination and intelligence are due to the intellectual factors in the tests.¹ Burt (16) finds no general connection between the capacity to discriminate lifted weights and intelligence as estimated by the school masters. The results of Thorndike, Lay and Dean (71) on 37 normal school women and 25 high school boys show low correlations between accuracy in reproducing lengths and intelligence as estimated by the pupils (25), by the teachers (12) and by the school records (—01); and between ability in weighing boxes to standards and intelligence as estimated by the pupils (23), by the teachers (08) and by the school records (21). From Thorndike's results, Simpson (58) estimates the probable correlation between general sensory discrimination and general intelligence to be about .23. In the light of the results of this investigation and those of other investigators it is safe to conclude that the test of arranging five weights has no diagnostic value and should be eliminated from the scale. If the test were changed so as to rule out the factor of sensory discrimination, and involve only the intellectual factors of comprehending a serial arrangement and making the logically necessary comparisons, the test would probably prove very valuable in differentiating the intelligence of younger children.

Constructing a Sentence. The test of constructing a sentence containing three given words with the resulting expression containing one or two ideas shows diagnostic values of 23% (1 idea) and 29% (2 ideas). In a preceding chapter (see page 129) it was seen that large differences existed in the character of the sentences constructed. The sentences "Trenton has lots of money on the river" and "Trenton paid a large sum of money to have the Delaware river deepened" represent widely different logical constructions, yet both would have to be scored plus under the "twelve year" test, for according to Binet the fact that a child constructs a single sentence containing the three words proves that he has a "mental age" of twelve, even if the sentence given be devoid of sense. An example given by Binet of such a sentence that would receive a "twelve year" credit is "Paris is a city of fortune by a stream." The writer believes that if the test were scored according to logical rather than grammatical merit, the diagnostic value would be much higher. Meumann (46) in an extensive study of the ability of children to construct sentences from two words, three words, and from pairs of words, places the greatest emphasis in the analysis of the results in their relation to intellectual endowment on the character of the responses.

Naming 60 Words in 3 Minutes. This test shows a Maximum Diagnostic Value of 40% when the passing mark is 75 rather than 60 words. In this test, 21% of the retarded group exceed the median of the normal group, 32% exceed the lowest 14 of the normal group and 84% exceed the lowest one.

¹ This applies to other tests more elementary than those of sense discrimination. Some of the correlations between intelligence and vital capacity are undoubtedly due to the fact that there is a trick in blowing up a spirometer, and that dull and defective children loose a lot of "wind." These correlations refer more to the quickness in sizing up the apparatus and catching on to the method than to the cubic contents of the mouth and lungs.

Giving the Day and Date. This test which in a preceding section was shown to depend on school training shows a diagnostic value of 29%.

Solving Problems from Various Facts. This test shows a diagnostic value of 35%, but this is not a true expression of the merit of the test for it is the resultant score of an effective and an ineffective test. Problem a (Hanging from a limb) shows a low diagnostic value (21%), while the other problem (Neighbor's visitors) shows a value of 51%. The second test is twice as effective as the first yet its merit is obscured by the scoring of the test according to the ruling that the subject must solve both problems in order to pass the test. The fact that one poor test may in this way lower the effectiveness of another test illustrates one of the advantages of the partial credit system adopted by Yerkes in the Point Scale. The explanation of the fact that the first problem shows a low diagnostic value probably lies in the fact that a large number of the normal group gave the answers "a bear," "a snake," etc., answers which to intelligent subjects seemed to be perfectly rational, but which had to be scored minus according to Binet's rule that "a man hanging" is the only acceptable answer.

Copying Designs from Memory. This test shows a Maximum Diagnostic Value of 36%. 9% of the retarded group exceed the median of the normal group, 52% exceed the lowest 14 and 88% exceed the lowest one. The fact that 33% of the normal 12, 13 and 14 year boys fail this test for "ten year" mentality would show that it is not a real "ten year" test. In all probability the visual memory involved is of a particular sort so that no group of individuals randomly selected would ever succeed in 100% of the cases.

Enumerating the Months. This test shows a diagnostic value of 43%. In a previous section it was shown that this test depended on school training. The examination of Chotzen's (18) results showed that they did not prove that the tests that showed the greatest increase with maturity were least dependent on intelligence. A test that depends on training may have a high diagnostic value, but the previous training of the subject must be known of course for the subject's failure to have significance.

Defining Abstract Terms. This test shows a diagnostic value of 51%. It is interesting to see that in this test the diagnostic value obtained by scoring a subject passed if he defines two or three terms correctly is higher than the diagnostic value of any of the three parts taken singly. The three words are equally difficult to define yet the word "kindness" has the smallest diagnostic value. That the test of defining abstract terms is too difficult is shown by the fact that only 59% of the normal subjects passed it, when the proportion should be approximately 90%. This test is one of the six that Binet considered as diagnostic of the mental differences between morons and normals.

Defining in Terms Superior to Use. This test shows a diagnostic value of 51%. All of the words are of the same difficulty except "mother" which is slightly more difficult due to the occasional embarrassment reaction that is encountered. The fact that 16% of the normal 12, 13 and 14 year subjects fail this test is probably due to the amused attitude that some older subjects assume when given this test. The consequence is short and careless answers. It is certainly true that all of the normal subjects were able

to give definitions in terms superior to use. Since this test shows such a high diagnostic value it would probably be well to change it in some way so that all subjects would have the same "Aufgabe," and so that the test would not depend on the random interpretations of "What is a ——?"

Detecting Absurdities in Statements. This test shows a diagnostic value of 53%. Absurdity *e* shows the highest diagnostic value (62%) and absurdity *d* the lowest (23%). Absurdities *a*, *c* and *d* are the easiest, absurdity *b* the most difficult, with absurdity *e* between.

Reconstructing Dissected Sentences. This test shows a diagnostic value of 71%. This test was found to depend on school training, the effect being noticeable between the fourth and fifth grades. This test was passed by 3 of the 10 retarded boys in the fifth grade, by 13 of the 29 in the fourth grade and by 1 boy in the third grade. None of the normal subjects in the sixth, seventh or eighth grades failed the test. Two interpretations of this test are of course possible. The first is that the high diagnostic value shown is due entirely to grade training. The second is that the test is entirely dependent on intelligence, and that children who have not sufficient intelligence to pass the test never reach the fifth grade. The truth probably lies in the view that the test depends on both factors. It can not be training entirely for the groups had been in school the same length of time, and a larger proportion of fourth than fifth grade subjects passed the test.

In designing the dissected sentence test, Binet sought to detect the same abilities that were involved in the Ebbinghaus mutilated prose tests. The results of Stenquist, Thorndike and Trabue (60) on a completion test show a very marked increase in the performance of children in the fifth and sixth grades over those in the third and fourth grades, the sudden increase in performance which indicates school training appearing between the fourth and fifth school grades, where the influence of this factor appeared in the dissected sentence test. The results of Fraser, reported by Whipple (76) show a higher correlation between performance on the completion test and scholastic status than between performance on this test and chronological age. Ebbinghaus (25) believed that the completion test involved factors most intimately connected with intelligence. Simpson (58) found that a completion test differentiated his two groups almost completely. Wyatt's (81) results show a high correlation between intelligence as estimated by the teachers and performance on the completion test (0.85, $p < 0.04$), a correlation higher than that obtained from any of the other 15 tests used. There seems to be good evidence then that both the Ebbinghaus completion test and its mutant offspring, the dissected sentence test, depend on school training and also correlate highly with intelligence.

Comprehending Difficult Questions. This test also shows a diagnostic value of 71%. Question *a* is practically impossible because the subjects did not understand the meaning of the word "delayed." The fact that the retarded children were 10% ahead of the normal children on this test may be due to chance or to the personal equation of the experimenter. Question *c* was passed by only 50% of the 12, 13 and 14 year normal children, showing that it is too difficult for "ten years." Questions *b*, *d* and *e* are of approximately the same difficulty, but questions *b* and *e* have a higher diagnostic value than

question *d* (76% and 74% to 51%). This test is included in the list of six tests that Binet considered to be valuable in differentiating the moron from the normal individual.

In the foregoing discussion of the 23 Binet tests used in this investigation, it was not possible to draw any conclusions concerning the tests of comprehending easy questions, repeating 5 digits, naming the pieces of money and making change, because they were too easy for both groups. It was found that the other tests varied in their diagnostic value from -71% to $+2\%$, or in other words it was shown that the scale contained some tests that were very effective and others that were quite ineffective. To summarize the results it is best to classify the tests according to the mental processes involved, in order to determine what sort of tests correlate best with intelligence.

Any classification of the tests according to the mental processes involved is of course inadequate, for these processes can not be determined except by experiment. The fact that a test is classified as involving a certain process does not prove that that process is involved. In fact two subjects may use quite different mental processes in solving the same test. The classification given in table 9 is offered with these qualifications. For the most part, the writer has adopted the analysis given by Yerkes (table 1, pages 7 and 8). The writer has added the factor of school training to the dissected sentence test, and of logical memory to the test of repeating sentences. The tests of solving problems, rhyming, naming the months and giving the date are not included in Yerkes' list, and were classified by the writer. The list of tests arranged according to their diagnostic value with the factors involved in each is shown in table 9.

In table 10 the tests have been re-classified according to the main factors involved in each, and these factors arranged in the order of their apparent worth in diagnosing intelligence. The diagnostic values of each test and of each part of each test are shown. All these values except those given for the tests of rhyming and repeating sentences are taken from table 6.

From table 10 it will be seen that most of the large differences between the normal and retarded groups appear under the head-

TABLE 9.

Factors Involved in the Various Tests.

Diagnostic Value	Tests
—71	Comprehending difficult questions. PRACTICAL JUDGMENT involving memory and imagination.
—71	Reconstructing dissected sentences. IDEATION involving analysis, imagination, command of language forms, school training.
—53	Detecting absurdities in statements. LOGICAL JUDGMENT based on imagination, analysis and reasoning.
—51	Defining terms superior to use. IDEATION (association and analysis).
—51	Defining abstract terms. IDEATION involving vocabulary.
—51	Problem b (Neighbor's visitors). PRACTICAL JUDGMENT, reasoning inductively from a concrete situation.
—43	Enumerating the months. SCHOOL TRAINING, memory.
—40	Naming 60 words in three minutes. ASSOCIATION (free), vocabulary, attention.
—36	Copying designs from memory. VISUAL MEMORY, perception, attention, motor coordination.
—33	Rhyming words with "defender" and with "day," "mill" and "spring." ASSOCIATION (controlled), vocabulary, attention.
—30	Repeating 7 digits. AUDITORY MEMORY for words (digits), attention.
—29	Giving the day and date. SCHOOL TRAINING, memory.
—29 and —23	Using 3 words in a sentence containing either one or two ideas. IMAGINATION and command of language forms.
—23	Repeating sentences (18 syllables). AUDITORY MEMORY for sentences, logical memory, attention.
—21	Problem a (Hanging from a limb). PRACTICAL JUDGMENT, reasoning inductively from a concrete situation.
—18	Arranging five weights. KINAESTHETIC DISCRIMINATION, ideation (notion of series), attention.
—12	Resisting suggestion. SUGGESTIBILITY, visual perception, comparison.
—6 and +2	Interpreting and describing pictures. PERCEPTION (visual—of things, relations, meanings), apperception, association, imagination.

TABLE 10.

Diagnostic Value of Various Tests and Parts of Tests Classified according to the Factors Involved.

	Diagnostic Value					
	total	a(1)	b(2)	c(3)	d	e
IDEATION						
Reconstructing dissected sentences*.....	-71	-70	-54	-65		
Defining in terms superior to use.....	-51	-49	-38	-42	-53	-43
Defining abstract terms.....	-51	-41	-42	-24		
JUDGMENT (logical and practical)						
Comprehending difficult questions.....	-71	+10	-76	-36	-51	-74
Detecting absurdities in statements.....	-53	-42	-46	-41	-23	-62
Solving problems from various facts....	-51	-21	-35			
SCHOOL TRAINING						
Enumerating the months.....	-43					
Giving the day and date.....	-29					
ASSOCIATION (free and controlled)						
Naming 60 words in 3 minutes.....	-40					
Giving rhymes with "defender".....	-33					
MEMORY (auditory and visual)						
Copying designs from memory.....	-36					
Repeating 7 digits.....	-30					
Repeating sentence (18 syllables).....	-23					
IMAGINATION						
Using three words in a sentence (2 ideas)	-29					
Using three words in a sentence (1 idea)	-23					
KINAESTHETIC DISCRIMINATION						
Arranging five weights.....	-18					
SUGGESTIBILITY						
Resisting suggestion	-12					
PERCEPTION						
Describing pictures	+ 2	- 3	+ 4	+ 4		
Interpreting pictures	- 6					
Emotional interpretation		0	- 2	- 3		
Intellectual interpretation		- 9	- 7	+ 6		

*The test of reconstructing dissected sentences also involves school training.

ings "Ideation" and "Judgment." Five of the 29 values under these headings are 70% or over, 14 are over 50% and 22 are over 40%. Under the remaining headings only two tests show a diagnostic value of 40% or over.

Two interpretations of these results are possible: the first, that the factors involved in the first six tests are those that are most intimately associated with intelligence; the second, that these 6

tests all involve the use of language, and that they are really diagnostic of the amount of linguistic training rather than intelligence. The 6 tests that show the highest diagnostic value most certainly deal with verbal material. One interpretation must hold that the normals handle the material better because they are more intelligent, the other that they handle it better because they have had more linguistic training. The position that the differences are due to training is supported by the fact that the dissected sentence test depends on school training, and that the test of enumerating the months also shows a high diagnostic value (43%). Against this position it may be said that the retarded children had been in school as long as the normal children, and had come from the same sort of homes. When the results of children of different linguistic training are compared, the differences are slight compared to those found between children of different intelligence. The normal children of English speaking parents average 5% higher than those of non-English speaking parents on these six tests, and the same difference is found between the retarded groups of different linguistic training. The normal group however averages 47% higher than the retarded group on these six tests. It is therefore legitimate to conclude that these tests involve other factors beside language—factors that are intimately connected with intelligence.

The reader may draw his own conclusions concerning the nature of the mental processes involved in the tests. In the writer's opinion, the tests that show the highest correlation with intelligence are those that involve reasoning—that demand the application of the subject's knowledge in a new way. It is safe to assume that a group of impartial judges would classify the mental processes involved in the six tests that show the highest diagnostic value among those commonly called the "higher thought processes," and would place the other processes of memory, imagination, sensory discrimination and suggestibility somewhat below these on a scale of complexity.

It is not necessary, however, for the purposes of this discussion to classify the mental processes involved in the tests. It is only necessary to note that certain tests show a very high diag-

nostic value while others show practically none. The tests of arranging weights, resisting suggestion and describing and interpreting pictures should either be changed so as to bring out their diagnostic value or thrown out of the scale entirely. The rest of the tests should be weighted in the scoring system according to their relative diagnostic value.

It has been seen that the Binet scale is inadequate in diagnosing the higher grades of mental defect. The reason for this is now obvious. Certain tests are diagnostic of intelligence while others are not. The subject receives the same credit for passing the tests that are not diagnostic as he does for those that are highly diagnostic. The subject who arranges the weights correctly receives the same credit (one fifth of a year) as the subject who answers all five of the comprehension questions or all five of the absurdity questions. In the same way a subject gets the same credit for passing the suggestion test as he does for defining abstract terms or reconstructing dissected sentences. A subject may pass all the tests of low diagnostic value and fail all the diagnostic tests and yet have the same "mental age" as a subject who passes the diagnostic tests and fails the others. For example, one subject may pass all the tests in VIII, all of the tests in IX except the definitions test, the weights, design and sentence tests in X, and the suggestion, sentence and 60 word tests in XII. He then has a basal age of 8, and having passed 10 tests above that, his "mental age" is 10. Another subject passes all of VIII and IX, the absurdity, comprehension and sentence test in X, and the definitions and dissected sentence test in XII. He then has a basal age of 9, and having passed 5 tests above that, his "mental age" is 10. The two subjects would have the same "mental age" yet the second would be far more intelligent than the first, because he passed all of the diagnostic tests while the other passed none of them. In some cases then the value of the diagnostic tests may be obscured by those that are not diagnostic. Such cases would of course be exceptional for according to chance normal children are just as apt to pass the tests that are not diagnostic as feeble-minded children. The fact that such cases are possible however certainly justifies the opinion of the

Buffalo conference (15) that "A mental age of 10 or above is not necessarily indicative of feeble-mindedness, regardless of how old the examinee may be."

The elimination of some of the tests of low diagnostic value would make the scale easier to apply and more accurate. This may be shown by comparing the total scores on the whole series with those on a few tests. There are two ways of scoring subjects on the Binet tests, by taking the "mental ages," and by subtracting these ages from the chronological ages and finding the age differences. The computation of "mental ages" over 10 is made difficult on account of the missing groups of XI, XIII and XIV. A child making basal 10 and passing 4 tests in XII has a "mental age" of $10\frac{4}{5}$. If he passed the additional test in XII his "mental age" would be 12, so that one test counts for a year and a fifth. In the same way a child passing all tests in XII and 4 in XV has a "mental age" of $12\frac{4}{5}$, but if he passes the additional test in XV his "mental age" would be 15, the extra test in this case counting for two years and a fifth. This difficulty may be overcome by weighting the tests in XII and XV, counting all tests in XII as two fifths of a year, and all tests in XV as three fifths of a year. The first case cited would then have a "mental age" of $11\frac{3}{5}$ which would become 12 if the additional test were passed, and the second case would have a "mental age" of $14\frac{3}{5}$ which would become 15 if the additional test were passed. The writer has scored the subjects in both ways, according to the conventional method and according to the method of weighting the advanced tests. The comparison of the "mental ages" with the chronological ages will also yield two measures depending on whether the conventional or weighted "mental age" is used. In treating the total scores, the writer has computed them according to all four methods in order that the most favorable method may be used for the purposes of comparison.

The scores of the subjects on five tests were used to compare with the total scores of the whole series. The scores of the subjects were computed for all five parts of the tests of defining in terms superior to use, all five of the absurdity questions, the last four comprehension questions, all three abstract definitions and

all three dissected sentences. The score was taken therefore on 20 parts of five questions so that the ability of each subject may be expressed anywhere on a scale from 0 to 20 points.

To compare the five methods of scoring (the "mental ages" and age differences according to the conventional and weighted methods of scoring, and the score on five tests), the medians of the normal and retarded subjects were found, the Maximum Diagnostic Value was computed, and the per cent. of retarded subjects exceeding the median of the normal subjects, the lowest 14 of the normal subjects, and the lowest normal subject was determined. These values are shown in table II.

TABLE II.
Comparison of Various Methods of Scoring the Binet Tests.

	Median of normal group (Q)	Median of retarded group (Q)	Per cent of retarded group exceeding the			Maximum Diagnostic Value
			Median of normal group	Lowest 14 of normal group	Lowest 1 of normal group	
Mental ages.....	10.8(0.2)	9.6(0.8)	2%	3%	41%	69%
Weighted mental ages..	12.0(0.5)	10.0(0.9)	3%	14%	41%	64%
Age differences.....	-2.6(0.75)	-4.0(0.7)	10%	22%	63%	52%
Weighted age difference	-1.3(0.9)	-3.6(0.8)	7%	22%	49%	54%
Score on 5 tests.....	16(2.75)	6(2.5)	3%	7%	19%	83%

From table II it may be seen that the effect of weighting the "mental ages" increases the quantitative differences between the medians of the groups, or extends the distribution of the measures. The difference between the conventional "mental ages" 1.4 yrs. (10.8 to 9.6) is increased to 2 yrs. (12.0 to 10.0), and the difference between the age differences is increased from 1.4 to 2.3. The variability of the measures is of course raised in each instance. The dispersion of the measures has however no effect on their effectiveness in differentiating the groups. The conventional method of scoring the "mental ages" shows the greatest differentiation between the groups and the greatest Maximum Diagnostic Value (69%). The score of the subjects on five tests is more effective in differentiating the groups however and

shows a higher diagnostic value (83%). In all, 23 Binet tests were used. These results show the paradoxical situation that the scale would have been more effective if 18 of these tests had not been given. This result is valuable only for a demonstration, and does not prove that only five tests should be used in testing intelligence, for the accuracy of any measure of general intelligence increases with the number of tests, provided that the tests are all effective. The result does show however that the method of increasing the accuracy of the measures of higher degrees of mental defect is that of increasing the number of diagnostic tests and eliminating the tests that are not diagnostic.

VI. DIAGNOSTIC VALUE OF SUPPLEMENTARY TESTS

The preceding discussion has been confined to the Binet tests entirely. In order to make the investigation more complete a number of other tests were given. The principle used in selecting the supplementary tests was that of diversity of character. Ten different sorts of tests were used, eight of them being apparently independent of language training. In cases where standard tests were used the procedure was adapted to the needs of the experiment. Deviations from standard procedure are to be excused on the grounds that age norms were not being sought, and to find the diagnostic value it was only necessary that the subjects understood the nature of the task they were to perform, and that the instructions were uniform for both groups. The detailed account of the method is given in the description of each test.

TEST I. PUZZLE TESTS.

A series of ten puzzles tests was used. The apparatus of one test was changed during the course of the experiment so the data are given on but nine. The two bicycle bell puzzles were given to all of the normal group and to 97% of the retarded group. All of the other puzzles were given all of the possible number of times. In a previous section, the value of these puzzle tests in obtaining the cooperation of the subjects was noted (see page 124). The discussion of their diagnostic value follows.

HEALY CONSTRUCTION PUZZLE A.

The material used for this test was the standard apparatus manufactured by C. H. Stoelting Co., and described under Test III on pages 14 and 15 of Healy and Fernald's monograph (34), and on pages 93 to 96 of Schmitt's monograph (57). The pieces of the puzzle were disposed irregularly on the table and the subject told to "Put this puzzle together." Healy's method of scoring is to record the number of moves, the number of

obvious impossibilities, the repetition of such obvious impossibilities, and the time, the subject being marked as failed if the time exceeded 10 minutes. In the present experiment, the number of moves and the time were taken, the subject being marked as failed if he made over 30 moves or did not succeed in solving the puzzle in a minute and a half.

81% of the normal subjects and 63% of the retarded subjects succeeded in solving the test under these requirements. The normal subjects solved the puzzle in from 5 to 27 moves, the median of all normal subjects (including those who failed) being 11.5 ($Q=7$). The median of the retarded group was 15 moves, the variability being higher than 10.5 moves. The Maximum Diagnostic Value for the test was 18% at the arbitrary passing limit selected (30 moves).

Schmitt classifies the responses to this test under three types—planned, trial and error, and chance. The subject is considered to have done the test by the planned method if he solves the test with less than 6 errors, by the trial and error method if he makes from 6 to 11 errors, and by the chance method if he makes more than 12 errors. The method of scoring used in this investigation may be compared to that used by Schmitt by adding the smallest possible number of moves (5) to the number of errors, and in this way considering the difference between the smallest possible number of moves and the actual number of moves the number of errors. A subject would be considered to have done the test by the planned method if he made less than 11 moves, by the trial and error method if he made from 11 to 16 moves, and by the chance method if he made 17 moves or over. 45% of the normal subjects and 36% of the retarded subjects did the test by the planned method, 19% of both groups by the trial and error method, and the remainder (36% of the normals and 46% retarded) by the chance method. This method of scoring therefore does not show a diagnostic value higher than 10%.

HEALY CONSTRUCTION PUZZLE B.

The material used for this test was the standard apparatus manufactured by C. H. Stoelting Co., and described under Test

IV on pages 16 and 17 of Healy's monograph, and on pages 97 to 100 of Schmitt's monograph. In this investigation, the number of moves and the time were recorded, the subject being marked as failed if he made 35 or more moves, or did not succeed in solving the puzzle in three minutes.

81% of the normal subjects and 51% of the retarded subjects succeeded in solving the test under these requirements. The normal subjects made from 11 to 34 moves, the median being 19 ($Q=9$). The retarded group solved the puzzle in from 11 to 34 moves, the median being 34 (the other half failing). The maximum diagnostic value was 35% if the passing mark were 30 moves.

Schmitt considers the subject as solving the test according to the planned method if he makes 8 errors or less, by the trial and error methods if he makes from 9 to 16 errors, and by the chance method if he makes 28 or more errors. The smallest possible number of moves is 11. Comparing the two methods of scoring, the subjects of this investigation who performed the test in 19 moves or less were classified under the planned method, from 20 to 27 moves under the trial and error method, and in 28 or more moves under the chance method. 50% of the normal and 24% of the retarded group do the test by the planned method, 16% of the normal and 15% of the retarded by the trial and error method, and the remainder (34% normal and 61% retarded) by the chance method. This method of scoring therefore shows a diagnostic value of 26%.

The comparison of the results of this investigation with those of Schmitt can not be made, for what Schmitt considers an "error" may not be what the writer considers a "move." The method of scoring is adopted merely to throw light on the results of this investigation. The criticisms of these two Healy tests which the writer makes are not made against the tests given according to Healy's procedure, but according to the writer's procedure under the conditions of this experiment. At a matter of fact Healy considers the quantitative scoring of his performance tests of little importance as compared to the qualitative scoring—to the experimenter's judgment of the manner in which

the subject approaches the problems. Schmitt has of course converted a quantitative scoring (number of errors) into a qualitative scoring (planned, trial and error, and chance methods), so that her results can not be used to criticize Healy's method.

Schmitt's results however taken on their own merit show one rather striking feature. The results on these two tests, classified according to chronological ages in tables XVII and XX on pages 95 and 98, show a fairly steady gain in the excellence of performance from 5 to $11\frac{1}{2}$, and a decrease from $11\frac{1}{2}$ to $14\frac{1}{2}$. Taking the three age groups, $10\frac{1}{2}$ to $11\frac{1}{2}$, $11\frac{1}{2}$ to $12\frac{1}{2}$, and $12\frac{1}{2}$ to $14\frac{1}{2}$, 90% of the first group, 77% of the second and 60% of the third solve construction puzzle A according to the planned method, while 90% of the first group, 71% of the second and 65% of the third solve construction puzzle B by the planned method. The results show therefore an inverse correlation with age from $10\frac{1}{2}$ to $14\frac{1}{2}$. The meaning of this is not exactly clear. Taken in connection with the fairly low diagnostic value found for these two tests (18% and 35%) it would seem to indicate that these tests had little value in diagnosing the higher grades of mental defect.

Haines (32) classified 63 subjects age 12 to 18 into three groups according to their performance on the Binet scale and the Point Scale. 21 of the subjects were classified as high grade morons, 16 as showing doubtful mental defect, and 26 as showing no mental defect. He found the construction puzzle A to be valuable for distinguishing the not defective from the doubtful, and the high grade defective from the doubtful, but found that construction puzzle B showed no definite diagnostic value. These results are quite opposite to those of this investigation in which puzzle B shows a higher diagnostic value (35%) than puzzle A (18%).

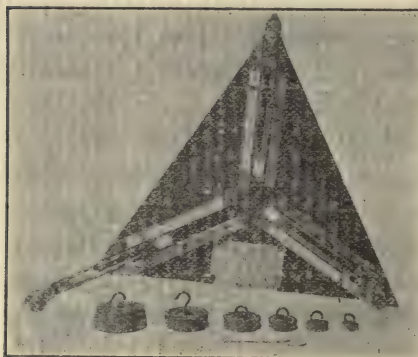
BICYCLE BELL PUZZLES A AND B.

Two mechanical puzzles were designed by the writer after the suggestion of Prof. H. C. McComas, bicycle bells being chosen to arouse the subject's interest. The various parts of the puzzles are shown in Fig. IA. Puzzle A consists of parts A, B, D, E and F. The position of parts D and E is shown inverted in B

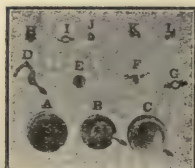
to show the manner in which the cogs mesh. Part F slips over the center pivot of B, to which the cover A is then attached, the completed puzzle being shown under C. The minimum number of moves of puzzle A is four. Puzzle B differs from puzzle A in that part F of puzzle A is broken up into parts G, H, I, J, K and L of puzzle B. In puzzle B parts D and E fit the same as in puzzle A. Part H slides on the center pivot, G fits over H, K and L fit in the ends of G and are held in place by I and the



D



C



A



B

FIG. 1

A. Bicycle Bell Puzzles.

B. Balance Test.

C. and D. Test of Lifting the Table Asymmetrically Balanced.

washer J. The minimum number of moves for puzzle B is eleven. The procedure in giving both puzzles was to place the parts in front of the child with the instructions "Put that bell together so it will ring." The subject was not allowed to see the bell as it was being taken apart. The time and the number of moves were recorded.

The subjects were considered as failing bell A if they took more than 20 moves or did not succeed in two minutes. 71% of the normal subjects and 47% of the retarded subjects passed the test. The median of the normal subjects was 15.5 moves, of the retarded subjects 27 moves. The Maximum Diagnostic Value was 24% when the passing mark was 20 moves.

Bell B was not given to the subjects who failed bell A, it being assumed that they would fail the more difficult one. Bell B proved entirely too difficult for both groups, being passed by but 22% of the normal subjects and 18% of the retarded subjects, the subjects being considered as failing if they took more than 30 moves or did not solve the puzzle in three minutes. The Maximum Diagnostic Value was 10% if the passing mark were fixed at 26 moves.

PUZZLES A, B, C, AC AND BC.

A series of puzzles was designed in order to test the subject's puzzle solving ability and his capacity to profit by experience. The puzzles are shown in Fig. 2 reduced one half. The puzzles were constructed of three-ply board painted white on one side and black on the other.

Puzzle A consisted of a triangular opening 9.8 cm. at the base, and 7.8 cm. on each side. Two right angle pieces $4.9 \times 6.05 \times 7.8$ cm. were provided to fill the opening. This puzzle is the same as the triangular portion of Healy's Introductory Picture Form Board described under Test I, pages 11 to 13 of his monograph.

Puzzle B consisted of the right angled piece fitting in the left side of puzzle A, and two other pieces, one $9.7 \times 6.05 \times 4.0$ cm., and the other $7.5 \times 4.9 \times 4.0$ cm. Puzzle B fitted in the same opening as puzzle A.

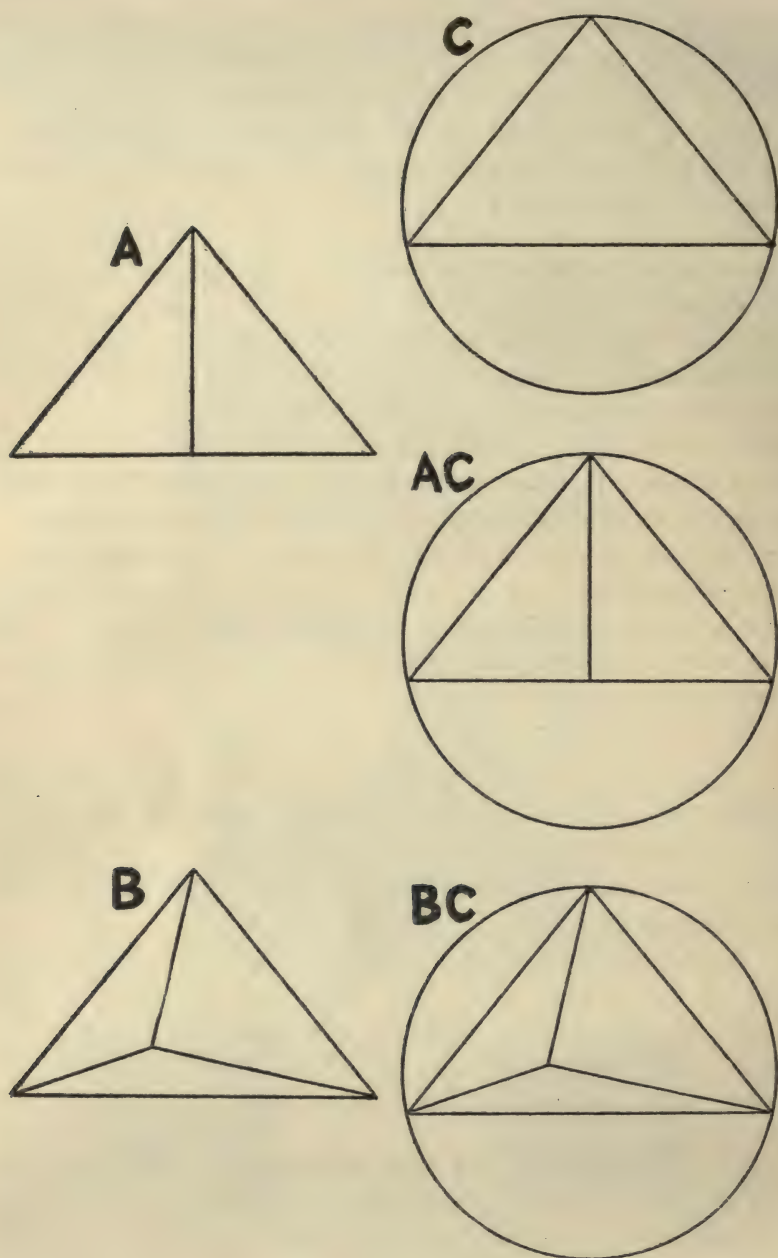


FIG. 2. Puzzle Series.

Puzzle C consisted of four pieces fitting into a circular opening 5 cm. in diameter. The triangular piece was the same size as the opening in puzzle A, 9.8 x 7.8 x 7.8 cm.

Puzzle AC consisted of the three curved portions of puzzle C and the two right angled pieces of puzzle A. In this test the subject has an opportunity to carry over the learning from puzzle A.

Puzzle BC consisted of the three curved portions of puzzle C and the three pieces of puzzle B, so that the learning from the latter puzzle might be carried over.

The instructions in every case were "Put those pieces in there. Keep the white side up." All five of the puzzles were given all the possible number of times. Puzzle A may be solved in two moves, puzzle B in three, puzzle C in four, puzzle AC in five and puzzle BC in six. Puzzles A, B, AC and BC were considered failed if the subject made 30 moves or over, or failed to accomplish the task in a minute and a half.

88% of the normal and 80% of the retarded passed puzzle A. The normal group varied from 2 to 28 moves, the median being 10 ($Q=7$). The retarded group varied from 3 to 26 moves, the median being 12 ($Q=7.5$). The Maximum Diagnostic Value was 14% when the passing mark was 5, 6, 7, or 8 moves. Schmitt classified a subject's response as trial and error if he tried each piece in more than two positions before finding the right one, and as planned if he tried one or both pieces in less than two positions before finding the right one. The subjects of this study are therefore classified as doing the test by the planned method if they make 6 moves or under, and by the trial and error method if they make more than 6 moves. 33% of the normal and 19% of the retarded used the planned method. The diagnostic value of this method of scoring is therefore 14%.

79% of the normals and 59% of the retarded passed puzzle B. The normal group made from 5 to 26 moves, the median being 15.5 ($Q=7$). The retarded group made from 4 to 27 moves, the median being 18 and the variability higher than 9 moves. The Maximum Diagnostic Value was 20% when the passing mark was placed at 25 moves.

Puzzle C was considered failed if the subject made over 15 moves. All of the normal group and 93% of the retarded group passed this test. The median number of moves of the normal group was 4 ($Q=1$), while that of the retarded group was 5 ($Q=2$). The Maximum Diagnostic Value was found to be 20% when the passing mark was 7 moves.

81% of the normal group and 61% of the retarded group passed puzzle AC. The normal group made from 5 to 28 moves, the median being 11 ($Q=6.5$). The retarded subjects made from 6 to 29 moves, the median being 19 and the variability higher than 9 moves. The Maximum Diagnostic Value was 27% when the passing mark was placed at 10 moves.

79% of the normal group and 59% of the retarded group passed puzzle BC. The normal group varied from 6 to 27 moves, the median being 13 ($Q=8.5$). The retarded group varied from 6 to 28 moves, the median being 22 moves and the variability higher than 9 moves. The maximum Diagnostic Value was found to be 26% when the passing mark was 15 moves.

To study the effect of learning from puzzles A to AC and from B to BC, the difference between the number of moves on the first test and the number on the second minus the three additional moves from the curved pieces was used. The coefficients of learning for all subjects were calculated according to the formulae,—

Coefficient of learning from A to AC = $A - (AC - 3)$

Coefficient of learning from B to BC = $B - (BC - 3)$

A positive value would therefore indicate that the subject had profitted by the experience of the first test, a negative value that he had not.

In the comparison of the scores from A to AC, the normal subjects varied from +25 to -21, the median being 0 ($Q=7.5$). The retarded subjects varied from +22 to -27, the median being 0 ($Q=9.5$). The Maximum Diagnostic Value was 11%.

In the comparison of the scores from B to BC, the normal subjects varied from +27 to -23, the median being +2.5 ($Q=4$). The retarded subjects varied from +27 to -26, the median being 0 ($Q=9$). The Maximum Diagnostic Value was 13%.

Although the series of puzzles was designed so that the practise on the first two puzzles might carry over to the last two, the medians indicate that very little if any of the learning was carried over. In only one case (normal group B to BC, +2.5 moves) is any learning shown. In the other cases the medians are zero indicating no learning. This is probably due to the fact that there is something in the way of a "trick" or a "catch" in these puzzles.

SUMMARY OF PUZZLE TESTS

The distribution of the number of moves made by normal (N) and retarded (R) subjects in solving each of the 9 puzzles is shown in table 12.

A study of table 12 shows that the two groups do not differ greatly in the character of their responses to the puzzles. The distribution of the number of moves is very much the same, and the writer can not find any marked differentiation between the groups in the character of their responses. The normal group is of course ahead of the retarded group in general. The median of the normal group is 3.5 moves lower than that of the retarded group on the Healy construction puzzle A, 15 moves on construction puzzle B, 11.5 moves on Bell A, 2 moves on puzzle A, 2.5 moves on puzzle B, 1 move on puzzle C, 8 moves on puzzle AC and 9 moves on puzzle BC. The highest diagnostic value found was 35% (Healy construction puzzle B).

The low diagnostic values found for the puzzle tests are surprising in view of the fact that the results from such tests are so frequently stressed as bearing considerable weight in mental examinations. Puzzle tests are used very largely at Ellis Island and at numerous clinics where mental examinations are made on individuals who have difficulty in the use of English. The reason that puzzle tests are stressed in such cases may be that no other tests can be used. Certainly the results of this investigation do not justify the confidence that is generally placed in such tests in the diagnosis of the higher grades of mental defect.

It may be argued that the element of chance operates in individual cases and might obscure a real correlation. If a subject

TABLE 12.

Distribution of Responses (Number of Moves) of Subjects in Solving Puzzle Tests.

M. D. V.	Healy A 18%	Healy B 35%	Bell A 24%	Bell B 10%	A 14%	B 20%	C 20%	AC 27%	BC 26%
Number of moves	N R	N R	N R	N R	N R	N R	N R	N R	N R
2					2				
3					5 3				
4			1 5		4 1	1	20 21		
5	4 4		2 1		4 3	1 2	11 13	4	
6	5 6		1 1		4 4	3 3	9 5	6 1	3 3
7	3 1		4		1 1	5 3	5 4	5 1	5 4
8	8 3		3		4 4	2 1	1 4	4 3	3 1
9	4 5		3 2		2 3	4	1 2	6 5	7
10	2 2		3 2		4 5	2 3	1 2	2 2	1 3
11	3	3 1	3 4	1	3 3	2 4	1 2	3 3	6 5
12	2 4	5 1	3 2	1	2 4	2 4	1	1 5	2 3
13	2	3 2	4 4	1	2 3	4 1		1 2	5 1
14	3 2	4 3		1	2	2 3	1	3 2	2
15	1 2	8 1	2		3	2 1		1	1
16	2 1	3 1	1 1	1 1	2 2	2 1		2	2 1
17	2 1	1 1	5 2		3	1 2		1	1 1
18	1	2 2	2 2	2 4	1	3 2		3 3	2
19	2	2		3	3 1	4 1		1 1	1
20		1 1	4 1	1	1	2		3	3
21	2	3		1	1			1	1
22	1 1	3 1				1 1		1	5
23	1	1		1	1 2	2 1		1	2
24		1 3			1	1			
25	1	1		1	1			1 1	1
26	2	1		1	2	1		1	4
27	1	2				1			1
28		1			1			1 1	1
29		2		3				1	
30		2							
31		1							
32		1 1							
33		1 2							
34		2 3							
Failed	11 22	11 29	17 30	45 47	7 12	12 24	0 4	11 23	12 24

has any such thing as a general puzzle solving ability, it would be unlikely that chance would operate against him in a series of nine puzzles, so that a combined score on all nine puzzles would guarantee the presence of any such general factor, while treating the combined score in the same manner as a single score would show

if this general factor were correlated with intelligence. The combined score of each subject was obtained by adding the number of moves that he made on each test. In case the subject failed the Healy construction puzzle A, bell B, puzzle A, B, AC or BC, his score was counted as 30 moves. Failure was counted 35 moves on construction puzzle B, 20 moves on bell A, and 15 moves on puzzle C. If a subject solved all the puzzles in the fewest possible number of moves his score would be 49, while if he failed all puzzles his score would be 250.

The total number of moves made by normal subjects varied from 93 to 220, the median being 138.5 ($Q=22.5$). The total number of moves made by retarded subjects varied from 96 to 245, the median being 171 ($Q=28.5$). The Maximum Diagnostic Value was found to be 34% when the passing mark was 150 moves. Combining the scores in order to give weight to any general puzzle solving ability therefore does not help matters any.

The fact that the puzzles used in this investigation did not show a high diagnostic value of course does not prove that all puzzle tests have the same low value. The results apply to this investigation only. In the writer's opinion however puzzle tests receive rather more weight in mental examinations than they deserve. The writer is frankly skeptical. The puzzles used in this series were all scored quantitatively, and it may be true that when they are scored qualitatively, when they are used merely to afford the experienced experimenter an opportunity to observe the subject's behavior on a concrete problem, they may have a high diagnostic value. Even then the experimenter should get his experience from examining a large number of normal individuals, for the writer has seen perfectly normal adults make the most impossible mistakes in solving puzzles, apparently because they had a definite attitude toward puzzles in general as being something to which reasoning does not apply, and in the solution of which the most rational method is that of trial and error. Of course intelligent persons do not place round blocks in square holes and leave them there, but this sort of ability is already satisfactorily tested by the form board.

The foregoing statements are simply an expression of the writer's opinion, and as such may be entirely false. The actual results only prove that the puzzle tests as given and as scored show a diagnostic value for the groups studied of but 34%. It is possible that if the shades of response scored were grosser the puzzles would be diagnostic of lower degrees of mental defect. The value obtained as an index of a test's merit by raising or lowering the difficulty of the test (the Maximum Diagnostic Value) is an accurate expression of the merit of a test only if the scoring is an adequate expression of the intellectual factors involved. As a matter of fact the puzzles used were well within the ability of the groups. The per cent. of subjects passing each test under the requirements is as follows,—

	Healy		Bell		Puzzles				
	A	B	A	B	A	B	C	AC	BC
Normal	81	81	71	22	88	79	100	81	79
Retarded	63	51	47	18	80	59	93	61	59

Healy says that most of his twelve year children solve construction puzzle A in from 12 seconds to 2 minutes, and he considers a subject as failing if he takes more than 10 minutes. The difference between the two minute response and the ten minute response cannot be expressed on a scale of 30 moves. It is quite possible that the puzzle tests might be diagnostic of the lower grades of defect if they were scored on a gross scale such as the number of minutes taken for the solution rather than the number of moves taken in a minute and a half. Under the conditions of the experiment it is not possible to determine the validity of this statement. It is only possible to show that the puzzle tests used have little value in diagnosing the higher degrees of mental defect.

TEST II. LIFTING THE TABLE ASYMMETRICALLY BALANCED.

A test was designed for testing practical judgment without the use of language in which the subject was required to lift with a hook a small triangular table asymmetrically balanced. A piece of board 1.4 cm. thick was cut in the form of an isosceles triangle 50.5 cm. on a side. 157 hooks numbered in rotation from left to right were attached to the upper surface of the board. On the bottom surface, legs 10 cm. long were attached to slides which could

be moved outward a distance of 12 cm., the distance being read on a millimeter scale. At the end of each slide a hook was fixed so that different sized weights could be suspended. The six weights used were ordinary 2 lb., 1 lb., 8 oz., 4 oz., 2 oz., and 1 oz. weights with hooks attached (the weights in their final form weighing 912 gms., 467 gms., 238 gms., 124 gms., 62 gms., and 33 gms. respectively). A button hook was furnished by which the subject was to lift the table. The apparatus is shown balanced in plate ID, the under surface being shown in plate IC.

By sliding the legs outward or inward and suspending different sized weights on the ends of the slides, the center of gravity of the apparatus could be thrown on any one of 43 hooks within a triangle whose outer surface ran about 13 cm. from the outer edge of the board. The table could be lifted by but one hook for each adjustment, one or two legs remaining on the floor for all of the remaining 156 hooks. 12 adjustments were used in this study, a scale for making them being attached to the under surface of the board. The table alone weighed about $1\frac{1}{2}$ kgms., and the heaviest possible combination of weights would make the whole apparatus slightly over 3 kgms.

The procedure in giving the test follows. The slides were adjusted and the weights attached while the subject was looking on. The experimenter then held the table a little above the floor and said, "You see this table is fixed with a heavy weight here and a little one here (etc.). I am going to give you a button hook, and you must pick up the table, so that it will balance, in just as few moves as you can. There is only one right hook to pick it up, and you must find that hook. If you get the wrong hook the table will tip up this way or that way (illustrating), if you get the right hook it will come up nice and level (illustrating). Now pick it up in just as few moves as you can." The table was placed on the floor, and the subject was not allowed to touch it with his hands. The instructions were given twice. The experimenter recorded the number of each hook in order as the subject tried them. No hint was given to the subject concerning the method but he was encouraged if he lost patience. The test was given to all the subjects, each subject having three trials.

The responses of the subjects varied from that of selecting the right hook immediately to that of selecting the hooks randomly and finally chancing upon the right hook. One subject made 97 attempts. Most of the subjects found the center of gravity in less than 10 moves. The distribution of the number of moves made by normal and retarded subjects in all three trials is as follows,—

No. of moves....	1	2	3	4	5	6	7	8	9	10	over 10
Normal	20	22	34	15	18	11	8	8	5	4	29
Retarded	20	18	20	17	13	12	9	5	10	5	47

The responses of the normal subjects varied from 1 to 42 moves, the median being 4 ($Q=2.5$), and those of the retarded subjects varied from 1 to 97 moves, the median being 5.5 ($Q=4$).

The factor of information was encountered and avoided by changing the adjustments. One 15 year subject made 23 attempts on the first trial and gave it up. On the second trial, after he had made 38 attempts the experimenter told him on which end of the board he could find the right hook, and he found it after 18 more attempts. On the third trial he found the center of gravity after 15 attempts. In spite of his poor performance he had enough intelligence to remember the number of the hooks by which he succeeded in lifting the table on his second and third trials, and he told the rest of the boys in his room. The experimenter used a different set of adjustments after this. The factor of information could enter by telling the subject to pick the table up near the end where the heaviest weight was suspended. There is no way of avoiding this factor.

The score of the subjects was not taken as the number of moves for the moves differed in their merit, 10 moves close to the right hook being a better response than 10 moves in various parts of the board. The distance in centimeters of each hook from each center of gravity was measured, and the score of the subject taken as the sum of the distances from each hook selected to the proper hook. For instance two subjects attempted to solve an adjustment which placed the center of gravity on hook no. 25. The first subject tried hooks 17, 33, 20, 15 and 25, the first four being 3 cm., 2 cm., 2 cm. and 5 cm. from no. 25, the solution. Another subject selected hooks 14, 17, 26, 29 and

25, these moves being 4 cm., 24 cm., 5 cm. and 3 cm. from no. 25. Both made five moves, but the total error of the first was 12 cm., and that of the second was 36 cm.

In the first trial the score of the normal subjects varied from 0 to 417 cm., the median being 31 cm. ($Q=35.5$ cm.). The retarded subjects varied from 0 to 1252 cm., to the median being 73 cm. ($Q=62$ cm.).

In the second trial the normal subjects varied from 0 to 371 cm., the median being 14 cm. ($Q=14.5$ cm.). The retarded subjects varied from 0 to 376 cm., the median being 15 cm. ($Q=23.5$ cm.).

In the third trial the normal subjects varied from 0 to 232 cm., the median being 7.5 cm. ($Q=9.5$ cm.). The retarded subjects varied from 0 to 1017 cm., the median being 17 cm. ($Q=21$ cm.).

The effect of practice is shown by the normal subjects in the reduction of the median from 31 to 14 to 7.5 cm. The retarded subjects improve from the first to the second trial (73 cm. to 15 cm.), but show no improvement from the second to the third trial (15 cm. to 17 cm.).

To study the diagnostic value the sum of the scores of each subject on each of the three trials was obtained. The sums of the scores of the normal subjects varied from 13 cm. to 798 cm., the median being 77 cm. ($Q=48$ cm.). The sums of the scores of the retarded subjects varied from 9 cm. to 2360 cm., the median being 121 cm. ($Q=83.5$ cm.). The Maximum Diagnostic Value was found to be 28% when the passing mark was 160 cm.

The test was given 131 times to 11 high school and college students. 14% of this group solved the problem in one move, 33% in two moves, 21% in three moves, 17% in four moves, 8% in five moves and 8% in more than five moves. The scores of these subjects varied from 0 to 94 cm., the median being 6 cm. ($Q=4.5$ cm.). The scores of the 12, 13 and 14 year normal subjects in their 174 trials, varied from 0 to 417 cm., the median being 15 cm. ($Q=16.5$ cm.). The maximum difference between the adult group and the normal group was found to be 35% when the passing mark was 15 cm.

The low diagnostic value found for this test (28%) indicates that it has little worth in differentiating the higher types of defect. It may be possible that the test would be serviceable in detecting grosser differences. 83% of the normal group and 73% of the retarded group solved the test in 10 moves or less so that the test was well within the ability of the groups. If a grosser means of scoring was used the test might detect lower degrees of defect.

TEST III. BALANCE TEST.

In order to test reasoning without language, a test was designed in which the subject was required to arrange weights subliminally different by means of a balance. The balance used was modified from C. H. Stoelting Co.'s Army Prescription Balance (catalogue no. 240). The balance was cut down and mounted on a new base as shown in plate IB. The weights used were five 200 gm. Stoelting Universal Laboratory Weights (catalogue no. 445), reduced as follows,—

no. 1	no. 2	no. 3	no. 4	no. 5
199.85 gms.	198.75 gms.	197.79 gms.	196.90 gms.	195.87 gms.

The first test given was to arrange three weights, the second five weights. In giving the test the experimenter would pick up two weights and say "These weights look alike but one is heavier than the other. I can't tell the difference by lifting them for they are too close together. If I want to find out which is the heavier I hang them on here this way (hanging the weights on the balance). Now which one is heavier?" After the subject had designated the heavier, he was given weights nos. 1, 2 and 3 with the instructions "I want you to put these weights in a row with the heaviest one first, then the middle one and then the lightest one, weighing them on the balance to find out which one is the heaviest, which one is middle and which one lightest." For the five weight test, the subject was given all five weights and told to "Put these in a row with the heaviest one first, and then the next, and then the next, etc."

The test as it stands is not worth standardizing, but it showed several suggestions that may be worth following out. The ob-

jectionable features to the test were due to the fact that the subjects were constantly on the alert for other criteria to judge the weights than the mere position on the balance. In the first place, they would try to sense the difference, and it was possible to discriminate between some pairs, although it was unlikely that they could be arranged in order. The experimenter would correct the subjects if they tried this method and tell them to use the balance. Occasionally the subjects would watch the pin and the scale, but as the weights were so unlike that one member of any pair always rested on the base of the balance, his method gave no help. If the subjects used this method they were told that it did not help. The third method was that of "watching the bounce" as the subjects called it. If weight no. 5 were on one side, and weight no. 1 were placed on the other, it would fall with greater force than weight no. 4 would in the same position. The subjects were also told not to use this method.

Inasmuch as the test was new the experimenter did not properly understand it, and gave the retarded group more credit than they deserved. The solution of the three weight problem depended somewhat on the chance selection of the pairs by the subject. Most of the subjects would leave the heavier weight of the first pair on the balance and compare the third with this. If the subject first compared nos. 2 and 3, and then compared 1 and 2, the problem was easy. If he first compared 1 and 3, and then compared 1 and 2, the problem was more difficult because another comparison (2 and 3) was necessary.

The results of this test are given with the reservation that they are not absolutely accurate. 90% of the normal group and 51% of the retarded group arranged the three weights correctly, making the diagnostic value 39%. This value should be higher for many of the retarded group were given credit for arranging the weights correctly even if the method were wrong, i.e., if they placed the weights in the right order without making all the necessary comparisons. 83% of the normal group and 34% of the retarded group arranged the five weights correctly, making the diagnostic value of this test 49%.

The writer has given the five weight test to many intelligent

adults, college professors, college graduates etc., and has found that the test is much easier for 12, 13 and 14 year normal boys than for normal adults. The child sees but one method of arranging the weights, that of elimination. He will make all the comparisons possible, always leaving either the heaviest or the lightest weight on the balance (usually the heaviest). When he has compared this one with all the others, he will sometimes try all of the other four again to make sure he has the heaviest one, and then place it to one side with the remark "There that one's heaviest" or "That one's king." He will then proceed through the remaining four in the same way, then through the three etc. His method is the longest one possible but it is absolutely certain. The intelligent adult not only tries to arrange the weights correctly, but he tries to do it in the fewest possible number of moves. He invents short cuts and tries to remember previous moves with the result that he frequently becomes lost in his own complications.

The three weight test is more valuable than the five weight test and more diagnostic if the conditions are controlled. The most important test is to give the subject nos. 1 and 3 for the first comparison, and nos. 1 and 2 for the second comparison, and see if he will make the arrangement without comparing 2 and 3. In the writer's experience intelligent children and adults will always refuse to make this unqualified generalization, while the unintelligent person is ever ready to accept it.

TEST IV. HEALY CROSS-LINE AND CODE TESTS.

The cross-line and code tests described under Tests IX, X and XI on pages 28 to 34 of Healy's monograph were given. The procedure was changed somewhat to meet the needs of the retarded group. The cross-line test A was used entirely as a practice test. A large "X" was drawn on the paper before the subject and the four numbers filled in. "You see I draw this cross and fill in the numbers. Now if I want to make a mark that stands for 1, I make it this way (drawing the symbol for 1), because the 1 here points that way. Now if I make a mark like this (drawing the symbol for 2), what would that stand for?"

In this way all four symbols were drawn and the subject practiced until he could name them all with the original diagram before him. The paper was then turned over and the subject required to name the proper digit for each of the marks as they were drawn before him in irregular order. If the subject named all four correctly, he was marked plus.

The cross-line test B was given with briefer instructions. The figure was drawn before the subject and digits filled in. The subject was asked to name the proper digit for the symbols of 7, 5 and 2. After he understood the arrangement he was told to study the figure carefully for the paper would be turned over and he would be asked to name all the figures. When the subject said he was ready, the paper was turned over and he was required to name all the symbols as they were drawn in irregular order. If he named any wrong, other symbols were given, and the wrong ones given later. If the subject named all nine digits correctly he was scored plus. The code test was drawn out for the subject in the same manner as the cross-line test and the system of dots explained. The subject was required to distinguish between the symbols for c and l, t and x. He was then told to study the figure, and when he said he was ready, the paper was turned and two or three symbols were given for each figure. The task was continued till the experimenter was certain that the subject could or could not perform the task. If the subject failed cross-line test B, he was marked failed on the code test without trying it. In this way the three tests were given to all the members of both groups.

98% of the normal subjects passed cross-line test A, 84% passed cross-line test B, and 57% passed the code test. 64% of the retarded subjects passed test A, 20% passed test B, and none passed the code test. The diagnostic value for the first test is therefore 34%, for test B 64%, and for the code test 57%.

Wyatt (81) found rather of a low correlation between the teachers' estimates of the intelligence of a group of 34 boys and girls and a cross-line test modelled somewhat after cross-line test B, the correlation (0.46, $p < 0.09$) standing twelfth in a list of fifteen tests. Wyatt's procedure probably accounts for the com-

paratively low correlation. After the cross-line test A had been drawn on the board of the class-room and explained, the second figure was exposed for 20 seconds and the subject required to fill in the symbols on a prepared blank. The short exposure probably made success in the test depend on visual memory rather than on the rational comprehension of the arrangement of the figure. This interpretation is supported by the fact that a test of memorizing letter squares correlated higher with the cross-line test than with any other.

Healy says of these tests "These three tests are especially noteworthy and valuable because their correct performance seems to demand mental powers which appear strongest in the normal adult mind and which are weakest in mentality of the child type" (Page 29). Concerning the cross-line test B he says "On account of the readily ascertained differences in performance between bright subjects and dull subjects, we have come to regard the test as extremely valuable" (Page 31). The results of this investigation certainly justify Healy's belief in the value of these tests. Only two of the Binet tests show a diagnostic value higher than 64%. Haines found the cross-line test B valuable in differentiating the not defective from the doubtful.

TEST V. MEMORY FOR COMMISSIONS.

Four tests of memory for commissions were used. The materials (a penny, key, knife, eraser, book, saucer, small card with 1c stamp, small card with 2c stamp, and two penny match boxes, one covered with blue and the other with red paper) were placed on the table in front of the subject. The following tests were used,—

Test 1. Give me the penny and then put the key in the saucer.

Test 2. Put the saucer on the book, then put the key in the saucer, and then give me the eraser.

Test 3. Put the eraser in the saucer, then put the penny on the book, then turn over the card with the red stamp on it, and then give me the blue box.

Test 4. Give me the red box, then put the eraser on the book with your left hand, then put the knife on the blue box, and then give me the card with the green stamp on it.

The commissions were read twice to the subject. During the reading the articles on the table were covered with a card board screen. Tests 1 and 2 were given to 95%, test 3 to 93% and test 4 to 75% of the retarded group. All four tests were given to all of the normal subjects. 100% of the normal subjects passed tests 1 and 2, 52% passed test 3, and 14% passed test 4. 98% of the retarded group passed test 1, 73% passed test 2, 22% test 3, and 2% test 4. The diagnostic value for the first test is therefore 2%, for the second 27%, for the third 30% and for the fourth 12%.

Inasmuch as the four tests are of the same sort they may be scored together on a scale of 10, each test being weighted according to its difficulty. 99% of all subjects passed the first test, 87% passed the second, 37% passed the third, and 9% the fourth. The tests were weighted on a scale of ten according to the formula,—

$$\frac{\text{reciprocal of per cent. passed}}{\text{sum of reciprocals}} = \frac{x}{10}$$

This formula gives the value of 0 for the first test, $\frac{1}{2}$ for the second test, 4 for the third test, and $5\frac{1}{2}$ for the fourth test. The scores of the normal and retarded subjects were distributed as follows,

No. of points....	10	6	$4\frac{1}{2}$	4	$\frac{1}{2}$	0
Normal	6	2	24	0	36	0
Retarded	0	1	10	2	30	13

Calculating the percentage passed by normal and retarded subjects at each passing mark from 0 to 10, the Maximum Diagnostic Value was found to be 35% when the passing mark was $4\frac{1}{2}$ points. This value is close to that found for the test of repeating digits (30%), and copying designs from memory (36%), but higher than that found for repeating sentences (23%). None of the memory tests shows a high diagnostic value.

Terman (65) reports "that after the age of 12 or 14 years memory for relatively meaningless material, like digits or nonsense syllables, improves but little; and that above this level it does not correlate very closely with intelligence (page 323).

Abelson (1) gave nine tests to 88 backward boys and 43 backward girls, averaging about 11 years in age, and found that a test of memory for commissions showed a higher correlation with "practical intelligence" than any of the other tests (.53 for girls and .65 for boys). However, he instructed the teachers to estimate the "practical intelligence" of the children by considering, in forming their opinion, which of the children they would soonest trust on an errand requiring the sharpest intellect. The correlation found possibly proves that a psychological test of ability to run errands correlates with the teacher's judgments of this ability, but proves nothing concerning the relation of this ability to intelligence.

TEST VI. DISTINGUISHING BETWEEN TERMS.

The subjects were asked to distinguish between three pairs of terms, "steam and smoke," "lie and mistake," and "laziness and idleness." The questions were asked in the form "What is the difference between—?" The answer required for the first pair, "steam and smoke," was that steam came from water while smoke did not. If the subject said that smoke was black and steam white, they were asked if they had never seen white smoke. The distinction between the second pair required was that a lie was intentional and a mistake accidental. It was not required that the words "accidental" and "intentional" be used, but that their meaning implied. For the third pair it was required that the subject imply that laziness was due to a subjective condition while idleness might be due to accidental circumstances.

98% of the normal group and 73% of the retarded group distinguished between "steam and smoke," making the diagnostic value 25%. 66% of the normal group and 27% of the retarded group distinguished between "a lie" and "a mistake," making the diagnostic value 39%. 45% of the normal subjects and 3% of the retarded subjects distinguished between "laziness and idleness," making the diagnostic value 42%. 85% of all subjects passed the first pair, 46% the second pair, and 24% the third pair. Weighting the tests roughly on a scale of 10 in the same manner as the commissions test was weighted, the first pair re-

ceives a value of 1 point, the second 4 points and the third 5 points. 1 retarded subject scored 10 points, 15 scored 5 points, 1 scored 4 points, 28 scored 1 point, and 14 scored nothing. 19 of the normal subjects scored 10 points, 7 scored 6 points, 18 scored 5 points, 1 scored 4 points, and 13 scored 1 point. Calculating the percentage passed at each passing mark, the Maximum Diagnostic Value was found to be 49% when the passing mark was 4 or 5 points.

TEST VII. SUBTRACTION TESTS.

In an article on the detection of higher grades of mental defect, W. E. Fernald (26) stated that feeble-minded individuals have difficulty in subtracting. A series of 10 subtraction tests was devised following the suggestion of Prof. H. C. McComas. On account of the time required by the test, only three of the tests were continued throughout the experiment. The three tests were as follows,—

Test 1. Subtract 3 from 16 till you get to 7, then subtract 2 till you get to 1.

Test 2. Subtract 4 from 30 till you get to 10, then subtract 3 till you get to 1.

Test 3. Subtract 3 from 43 till you get to 25, then subtract 4 till you get to 9, then subtract 2 till you get to 1.

A preliminary practice test was given in which the subject was asked to subtract such figures as 4 from 13, 4 from 31, 3 from 22, etc. If he could not do these the rest of the test was not given. The problems were repeated twice for the subjects.

98% of the normal group passed test 1, 90% passed test 2, and 60% passed test 3. 24% of the retarded passed test 1, 17% passed test 2, and 7% passed test 3. The diagnostic value for the first test is therefore 74%, for the second test 73%, and for the third test 53%. The retarded group had great difficulty in passing the first problem and even in doing any sort of work in subtraction. This ability would of course depend on school training, but if the amount of a child's training is known, the results show that this test would be very valuable in diagnosing intelligence. The retarded group on an average had been in

school over 7 years, yet only 24% of them passed the first test. The fact that the experimenter gave himself an opportunity of deciding on the basis of the preliminary problems whether or not the other problems should be tried, affords an opportunity for the influence of the personal equation. It is very possible that if the experimenter had exercised more patience in some cases more of the retarded group would have passed the easier tests, (nos. 1 and 2), and the diagnostic value been lowered. On this account the diagnostic values obtained (74% and 73%) are probably too high, but the fact remains that there is a very large difference between the normal and retarded groups in their performance on these tests. The writer has used simple subtraction problems in examining mature defectives and suspected cases of dementia, and has found the tests very helpful and suggestive. All individuals must do a certain amount of subtracting in their general daily experience in counting change etc. The general custom among store keepers of counting the change up from the amount of the purchase to the amount of the piece of money given is probably evidence that some of their customers have difficulty in subtracting.¹

TEST VIII. SUGGESTION BY PROGRESSIVE LINES.

A test of suggestion by progressive lines, modelled somewhat after Test 42 of Whipple's Manual (76)² was used. The apparatus consisted of 13 cards 4 x 28 cm., each card having a line drawn 2 cm. from the left side equidistant from the top and bottom. The lengths of the lines on the first six cards were 10, 20, 30, 40, 50 and 60 mm. The lines drawn on the last seven cards were all 60 mm. in length. The subject was given a sheet

¹ In giving the subtraction tests, the writer has met many newsboys who can not do problems such as 4 from 22, 5 from 31, etc., but who can make change from some of the smaller pieces of money by means of finger counting systems, etc. The concrete problems are much easier.

² The test described by Whipple involves the use of a kymograph which is inconvenient. The test was shortened by using 13 lines instead of 20. The coefficient of suggestibility used in this investigation is more accurate than that given by Whipple because the measure of a subject's ability to reproduce lengths is obtained from six reproductions of lengths instead of one.

of 2 mm. cross section paper with heavy rulings at each centimeter. The instructions were "I am going to show you cards like this, and this (showing card no. 1, and one of the last seven), and I want you to draw lines on the paper just as long as the ones I show you." The subject was shown where to draw the lines on the paper, beginning at the left hand margin. The subject was allowed to study each line as long as he wanted to, and as soon as he started to draw a line and turned his attention to the paper, the card was dropped and the next one exposed.

The suggestion arising from the fact that the first six lines increase in length will usually cause the subject to increase the length of the last seven lines. The measurement of the amount of suggestion is made by comparing the increase in length of the last seven lines with the measurement of the subject's ability to reproduce the first six lines. The total actual length of the first six lines is 210 mm. The measurement of the subject's ability to reproduce lengths is the difference between 210 mm. and the sum of the lengths drawn. If no suggestion were present the length of the last seven lines would be to the length of the first six as 420 mm. (the actual length of the last seven) is to 210 mm., or the sum of the last seven should be twice the sum of the first six. The measurement of the amount of suggestion is therefore the difference between the sum of the last seven lines and twice the sum of the first six.

The lines drawn by the subjects were measured within 2 mm. and the length of each line recorded. The sum of the first six (SF) and last seven (SL) lines were then computed and the coefficients of accuracy and suggestion computed according to formulae explained,—

$$\text{Accuracy in reproducing lengths} = \text{SF} - 210$$

$$\text{Coefficient of suggestion} = 2\text{SF} - \text{SL}$$

A coefficient of accuracy with a plus sign indicates that the subject over-estimated the lengths, a minus coefficient that he under-estimated the lengths. A minus coefficient of suggestion indicates the influence of suggestion while a positive coefficient indicates no influence or a negative influence.

In general the subjects under-estimated the length of the first six lines and accepted the suggestion of increasing lengths. One subject increased each line one centimeter so that the last line was 130 mm. in length, over twice the length of the line to be copied. The average length of each line was computed for normal and retarded subjects. These results are given in Table 13, and are shown graphically in Fig. 3.

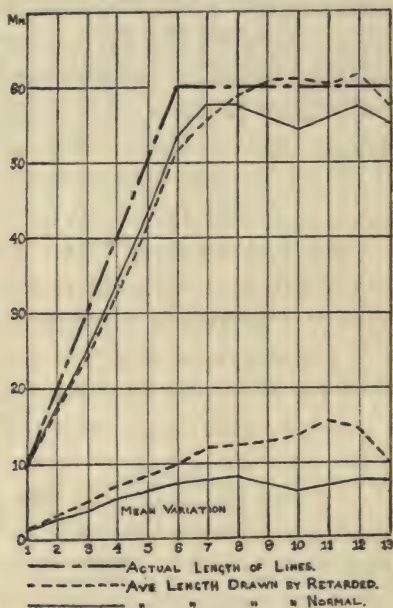


FIG. 3. Results of Normal and Retarded Subjects on Test of Suggestion by Progressive Lines

The normal subjects show the greatest influence of suggestion on the 7th and 8th lines, while the suggestion is still fresh. They then throw off the suggestion a little on the 9th line and more on the 10th but increase on the 11th and 12th, dropping back again on the 13th. The retarded subjects increase to the 9th and 10th, drop slightly on the 11th, increase on the 12th, and drop back on the 13th. The sum of the average lengths of the first six lines for normal subjects is 181.26 mm., and for retarded subjects 176.24 mm. The normal subjects underestimate

TABLE 13.

Suggestion by Progressive Lines.

Average Length and Mean Variation in Millimeters of Each Line Reproduced by Normal and Retarded Subjects.

Line no.	Actual length	Normal subjects	Retarded subjects
1	10	9.34 (1.12)	9.69 (1.29)
2	20	17.34 (2.59)	17.19 (3.10)
3	30	24.79 (3.79)	24.68 (4.88)
4	40	33.48 (5.38)	32.05 (6.85)
5	50	43.17 (6.19)	41.22 (8.34)
6	60	53.14 (7.31)	51.41 (9.83)
7	60	57.62 (7.69)	55.47 (12.00)
8	60	57.62 (7.95)	58.56 (12.20)
9	60	55.90 (7.21)	60.73 (12.56)
10	60	54.27 (6.31)	60.90 (13.52)
11	60	55.76 (6.93)	60.15 (15.44)
12	60	57.17 (7.66)	61.47 (14.78)
13	60	55.03 (7.61)	57.15 (9.95)

the length of the first six lines 28.74 mm., and the retarded subjects 33.76 mm. The sum of the average lengths of the last seven lines for normals is 393.37 mm., for retarded 414.43 mm. These totals compared to the ability to reproduce the first six lines show a coefficient of suggestion for normals of -30.84 mm., and for retarded of -61.95 mm. In general then retarded subjects underestimate the length of the first six lines and overestimate the length of the last seven lines more than normal subjects.

The coefficients of accuracy found for normal subjects varied from $+28$ mm. to -92 mm., the median being -30 mm. ($Q=23$ mm.). The coefficients of accuracy for retarded subjects varied from $+54$ mm. to -112 mm. the median being -24 mm. ($Q=29$ mm.). The Maximum Diagnostic Value was found to be 19% when the passing mark was taken at -50 mm.

The coefficients of suggestion found for normal subjects varied from $+50$ mm. to -210 mm., the median being -26 mm. ($Q=22$ mm.). The coefficients of suggestion found for retarded subjects varied from $+86$ mm. to -342 mm., the median being -56 mm. ($Q=45$ mm.). The Maximum Diagnostic Value was found to be 28% when the passing mark was -74 mm.

The diagnostic value found for the Binet line suggestion test was 12%, and this value taken in connection with the maximum diagnostic value shown by this test (28%) is fairly conclusive evidence that the correlation between intelligence and suggestibility is not very high, and that tests of suggestion have little value for diagnostic purposes.

TEST IX. ESTIMATION OF LENGTHS.

A test of estimating lengths was arranged after the suggestion of Prof. H. C. McComas for obtaining a simple intellectual judgment on sensory material. Four lines were drawn on each of ten cards $12 \times 7\frac{1}{2}$ cm. The top line was in all cases 100 mm. in length. The other lines were of lengths such that two of them if combined would be equal in length to the standard line. Ten cards were used, graded in difficulty so that the error of the wrong combinations varied from 25 mm. to 5 mm. The lines were drawn equidistant from the sides of the cards. Two of the cards (nos. 1 and 7) are shown in Fig. 4.

In giving the test the cards were presented in order. The instructions were "Here is a card with a long line at the top and three other lines here. Two of these lines if put together will make the top line. Which two lines would you put together to make the top line?" After the subject had judged the first card, whether he answered right or wrong, the lines were measured

TABLE 14.

Estimation of Lengths

Length of Lines Judged and Percentage of Correct Judgments.

Card no.	Length in mm. of lines no.			Per cent. passed by normal	Per cent. passed by retarded
	1	2	3		
1	75	50	25	88	71
2	75	25	5	98	86
3	90	70	30	76	64
4	70	30	10	100	81
5	65	50	35	60	54
6	80	65	35	57	47
7	60	40	30	93	75
8	70	60	40	41	29
9	55	50	45	28	29
10	60	55	45	21	10

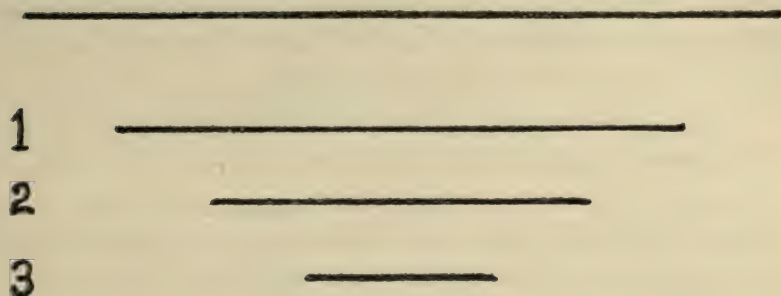
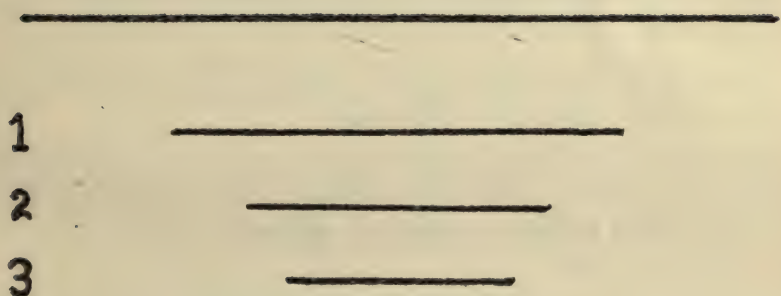
NO.1**NO.7**

FIG. 4. Estimation of Lengths Test.

off on a piece of paper, added together, and compared with the standard in order to illustrate the principle of the test. The subject was allowed all the time he desired to make the judgments. The measurements of the lines on each of the ten cards, and the per cent. of correct judgments on each card are shown in table 14.

From table 14 it may be seen that the difference between normal and retarded subjects is not more than 19% on any one set of lines. Scoring the test as a whole by counting the number of correct judgments in ten, the record of the normal and retarded subjects was as follows:

		no. of correct judgments									
		1	2	3	4	5	6	7	8	9	10
Normal		0	0	2	4	16	3	12	11	8	2
Retarded		1	0	3	12	14	16	6	6	1	0

Calculating the percentage that would have passed had the passing mark been fixed at any point, the Maximum Diagnostic Value was found to be 35% when the passing mark was 7 correct judgments in 10.

TEST X. CLASS-ROOM REASONING TEST.

A special test was designed to involve reasoning ability. The test included seven questions with two to four sub-questions under each. The questions varied greatly in difficulty so that the test could be given to a wide range of subjects. A special blank was prepared with questions printed on both sides of a paper 22 x 26 cm. Places were marked at the top of the test blank for the subject to fill in his name, school, grade, the date, and the date of birth. The directions printed at the top of the blank were as follows:

Write your name, school, grade, date of birth and date in the places marked at the top of the paper.

Answer each of the questions below in one or two words. Write your answer in the place marked "Answer" like this:

- a. When are days warmer, in February or June?

Answer—June.

- b. Is it because the sun's rays are slanting or straight at that time?

Answer—Straight.

All the questions may be answered in one or two words, such as "Yes," "No," "Right," "Left," "Not necessarily," "Wrong," etc.

Under each question a space was left after the word "Answer—." The questions asked were as follows:

- 1, a. When are days the longer, in January or in July?
b. Is it because the sun rises earlier or later?
- 2, a. If anything floats in water, is it lighter or heavier than water?
b. If a thing is heavier than water, will it float or sink?
c. If a thing floats in air, will it float in water?
d. If a thing sinks in air, will it sink in water, too?
- 3, a. Where is the sun in the morning, North, West, East or South?
b. If you face the South, will the West be on your right or your left?
c. If you face the North in the morning, which side will your shadow fall on, your right or your left?
- 4, If you put a stick in the water on a slant, it will look as if it were bent upwards.
a. If you were going to spear a fish from the side, would you aim above or below the fish?
b. Why? Is it because the fish seems to be above or below the place where he really is?
- 5, a. When are shadows longer, in summer or in winter?
b. Is it because the sun is farther North or farther South?
- 6, Supposing a clock starts running backwards at six o'clock.
a. In six hours, will it tell the right or the wrong time?
b. In twelve hours and a half, will it tell the right or the wrong time?
Supposing another clock starts running backward at half past six.
c. In six hours and a half, will it tell the right or the wrong time?
d. In twelve hours will it tell the right or the wrong time?
- 7, Water and cream both float on milk because they are lighter than milk. Water is heavier than cream.
a. Will cream float or sink in water?
b. If a thing floats on milk, will it float on water?
c. Supposing something sinks in cream, will it sink in milk?
d. Will a thing float on water if it floats on cream?

Questions 1a, 1b, 2a, and 5a were taken from questions 1, 4, 7 and 9 of set *a*, test IIIB of Bonser's (12) list of tests. The rest are new. The test was given individually and as a class room test. When given individually, the experimenter would give the subject a paper and read the instructions to him. The experimenter would then read each question and record the subject's answer, the subject in each case having a paper so that he might read and follow the experimenter. When given as a class room test, the teacher of the room would distribute the papers and tell the class to read the instructions and start to work. Any questions asked were referred to the instructions. As the children finished, they would bring their papers to the teacher who would glance over them to see that all the questions had been answered. If any questions were unanswered, the children were made to return to their seats and fill them in. No time limit was set in either the individual or group testing. The subjects were allowed all the time necessary. The test very rarely took over 20 minutes. The subjects were told that they might look at a clock or watch for question 6. The class-rooms had clocks on the walls, and the experimenter showed the individual examinees his watch.

The test was given individually to 56 members of the retarded group. It was given as a class-room test to all the children in the seventh and eighth grades of the Franklin School in Trenton, and to a class of college juniors and seniors in Princeton. The experiment at the Franklin School was conducted by the departmental teacher of geography who followed the instructions carefully. The experiment on college juniors and seniors was conducted by the professor in charge of the course with the assistance of the writer.

The age and grade distribution of the seventh and eighth grade girls is as follows:

	Age						Tot.
	11	12	13	14	15	16	
Grade VII	6	18	24	8	4	2	62
Grade VIII		2	15	19	15	1	52

The age in grade distribution of the seventh and eighth grade boys is as follows:

	Age					Tot.
	11	12	13	14	15	
Grade VII	2	22	24	21	5	74
Grade VIII		3	12	15	7	37

The average age (at last birthday) of the seventh grade girls is 12.87 yrs. ($MV=0.87$ yrs.), of the eighth grade girls 13.98 yrs. ($MV=0.67$ yrs.), of the seventh grade boys 13.07 yrs. ($MV=0.79$ yrs.), of the eighth grade boys 13.70 yrs. ($MV=0.74$ yrs.). The seventh and eighth grade boys included 53 members of the normal group.

The age distribution of the 41 college juniors and seniors was as follows:

Age	19	20	21	22	23	24	Tot.
No. of subjects..	1	11	12	14	2	1	41

The average age of the college students was 21.20 yrs. ($MV=0.86$ yrs.)

The results were calculated for each group on each question. In 18 of the 21 questions the right and the wrong answers were given in the questions. Questions 2d, 7b and 7c had only one answer possible, "Not necessarily" or some equivalent such as "Not always." "Sometimes" etc. Question 3a had four possible answers, so that 25% could answer correctly by chance. The remaining 17 questions could be passed by 50% by chance. The results were also calculated for combinations of parts of questions. One half of the subjects could pass either part of question 1, by chance, but only one fourth could answer both parts correctly. The scores were calculated for passing all parts of questions 1, 3, 4, 5, and 6, for parts a and b of question 2, for parts a, b, and c of questions 2 and for parts a and d of question 7. The chances are 1 out of 4 for passing 1ab, 2ab, 4ab, 5ab and 7ad, 1 out of 8 for passing 2abc, and 1 out of 16 for passing 3abc and 6abcd.

The results are shown in table 15 for each question and for the various combinations of questions as passed by the retarded and normal groups, by seventh and eighth grade boys and girls and by college adults.

If the personal equation figured in the results of this test it could only affect the retarded group as the test was given to all

TABLE 15.

Per Cent. that 322 Subjects Passed Reasoning Tests.

No. of subjects Test	Retarded	Normal	Grade VII	Grade VIII	Grade VII	Grade VIII	Adults
	group.	group.	boys.	boys.	girls.	girls.	
	56	53	74	37	62	52	41
1a	64	98	95	100	90	92	95
1b	71	89	88	89	85	85	88
1ab	46	89	85	89	82	83	88
2a	82	96	93	97	79	94	98
2b	96	98	95	100	84	92	98
2c	52	79	76	84	60	50	98
2d	7	17	15	16	11	29	93
2ab	80	96	92	97	74	92	98
2abc	46	75	70	81	47	46	95
3a	45	89	89	95	87	87	100
3b	48	72	70	81	50	63	98
3c	52	68	69	81	61	60	100
3abc	18	53	46	73	39	37	98
4a	41	55	66	62	52	50	83
4b	21	57	53	51	42	35	88
4ab	11	38	38	38	21	17	80
5a	16	57	53	51	40	56	78
5b	54	58	47	70	45	54	56
5ab	4	25	23	30	11	25	46
6a	27	72	74	81	68	75	95
6b	63	60	72	54	44	52	85
6c	54	49	45	46	37	44	78
6d	63	66	59	70	56	65	88
6abcd	5	30	27	27	6	19	66
7a	66	85	77	73	74	77	85
7b	0	8	4	5	6	8	76
7c	0	9	4	11	5	12	68
7d	66	91	76	97	63	75	76
7ad	48	75	59	73	45	65	66

other subjects in classes. If a question is too difficult for a group the percentage passed should at least approximate that expected by chance. According to chance 50% of all subjects should pass a question with two alternatives such as 5a. Table 15 shows that only 16% of the retarded group pass this test, so that it is fair to assume that this question was weighted against the subject in some way. Only 21% of the retarded subjects pass question 4b and only 27% of that group pass 6a, it being legitimate to expect 50% by chance. If there is anything in the form of the question, any popular misconception, or any other constant factor that would tend to make the wrong answer appear more frequently, the same question should show a deviation in the same

direction in the results of the other subjects. Minus deviations of 8% and 15% occur in the results of seventh and eighth grade girls on question 4b, and of 10% in question 5a. Minus deviations occur in the results of the groups to whom the test was given in class, but in no case are these deviations more than 15% below that expected by chance. It is right to assume then that questions 4b, 5a and 6a were influenced by the personal equation of the experimenter.

The test was given individually to the retarded group by the writer who is positive that there was no conscious intention to throw the results one way or another. However it seems legitimate to assume that there was something in the way the questions were read, a slight stress of the voice on the wrong alternative, a factor not consciously analyzable by the experimenter but one that was strong enough to throw the results in a definite direction in the long run. If this is not the case, and a minus deviation of 34% may still be attributed to chance, it is necessary to assume that a plus deviation of 34% may also be due to chance. Or again, if the experimenter unconsciously forced the results against the subjects in some questions, he might have favored the subject in others, and if 34% is the magnitude found in one direction, it is fair to expect an equal deviation in the opposite direction. The results of the retarded group are above 84% in only one question, 2b. It is therefore necessary to discard all the results of the retarded group on this test. No comparisons can be made with the normal group, and no indication obtained of the diagnostic value of these tests for the normal and retarded groups.

The comparison of seventh and eighth grade boys and girls shows slight sex differences in favor of the boys. Comparing the differences between the results of the sexes in both grades in all tests scored individually and in combination (58 comparisons), the median is 7.5% ($Q=6.5\%$) in favor of the boys. The differences are 20% or higher in 17% of the cases, and higher than 30% in 3% of the cases. The largest differences are in favor of the boys in question 2abc and 3abc.

The comparison of the results of seventh and eighth grade

boys and girls (range of ages from 11 to 16) with the results of college adults (range of ages 19 to 24) affords some indication of the value of these tests in differentiating adolescents from adults. Tests 7ad, 1ab, 2ab, 2abc and 3abc are too easy for adolescents to afford any diagnostic value. Test 5ab is apparently too difficult for adults and therefore worthless. Test 6abcd is rather hard for adults. The subnormal group might have guessed the answers to the clock questions, but the writer is certain on the basis of the individual examinations, that very few could figure the answers out. This ability seemed uniformly present in college adults, the mistakes being due to carelessness. Of the 164 questions in this test answered by college adults, 87% of them were answered correctly. Of the 900 questions in this test answered by the seventh and eighth grade boys and girls, 59% were answered correctly, only slightly better than the 50% to be expected by chance. If the question had been worded "What time would it be?" or the test arranged to bring out this factor, the diagnostic value of the test would probably have been demonstrated. Test 4ab seems to show considerable value in differentiating adolescents from adults. This test is passed by 28% of the seventh and eighth grade subjects, 25% being expected by chance, and is passed by 80% of the college adults.

The greatest differences between the adolescent and adult groups are shown in questions 2d, 7b and 7c, all of which involve the answer "not necessarily." The child almost always answers "yes" or "no" to these questions. It would seem from this that the child is willing to generalize too easily. The adult generally refuses to make an unqualified generalization. The child stumbles into it. This same factor appeared in the test of arranging three weights by means of a balance. Tests involving this factor would seem to be worth developing in the extension of measuring scales upwards.

It is interesting to note that Martin (44) finds this same type of response a valuable test for the upper years. Martin applied the Binet and DeSanctis tests to 212 normal children and 150 feeble-minded children. She noted the character of the responses of the subjects to questions 6a of the De Sanctis series

("Are large things heavier or lighter than small things?"), which was intended as a preliminary question to 6b ("How does it happen that sometimes small things are heavier than large things?"). It was found that children of the higher "mental ages" tended to qualify their statements by saying "that depends on the material," "large things are usually heavier," or by such words as "many," "sometimes," "often," etc.

Martin found that 24% of the 21 feeble-minded subjects of "mental age" 10 and 20% of the 10 normal subjects of the same "mental age" gave qualified answers. Only 2 of the 9 normal and feeble-minded subjects of "mental age" 11 failed to give qualified answers. The test was also given as a class-room test in several school grades. 27% of the children in the fourth grade, 68% of those in the fifth grade, and 81% of those in the sixth grade gave qualified answers. Martin concludes that "If note is made of the qualified answers, it would seem that the question is quite valuable in itself and might be used among the tests for the upper years." (page 102)

Inasmuch as questions 2d, 7b and 7c of the present investigation are obviously too difficult for seventh and eighth grade children, and Martin's data show that the qualifying response to question 6a of the De Sanctis series is well within the ability of the fifth and sixth grades, it would seem that the intellectual level is indicated by the refusal to make an unqualified generalization *from given material* rather than by anything in the nature of the process itself. The level is apparently indicated by the refusal to make a certain generalization, not by a change in the character of the reasoning process itself.

The form in which the whole reasoning test was given is unsatisfactory for individual testing. The method in which the right and the wrong answers are given is valuable for a class-room test for it saves time. The large number of subjects that may be obtained by the method shows whether an ability is present or absent. One can never tell however whether an individual child has not guessed the answer. There is no reason why the form of the questions should not be changed, and the tests used to bring out the factors found valuable in differentiating adolescents from adults.

SUMMARY OF SUPPLEMENTARY TESTS.

The list of supplementary tests used arranged in the order of their diagnostic value is shown in table 16.

TABLE 16.
Diagnostic Value of Supplementary Tests.

	% passed normal	% passed retarded	Diagnostic value
Subtraction Test no. 1*.....	98	24	74
“ “ no. 2*.....	90	17	73
“ “ no. 3*.....	60	7	53
Healy Cross-Line Test A.....	98	64	34
“ “ B.....	84	20	64
“ Code Test	57	0	57
Balance Test, 3 weights*.....	90	51	39
“ “ 5 weights*.....	83	34	49
Distinguishing between Steam and Smoke*.....	98	73	25
“ “ Lie and Mistake.....	66	27	39
“ “ Laziness and Idleness.....	45	3	42
Weighted score on all three pairs.....	(Max. diag. value)		49
Estimation of Lengths*.....	“	“	“ 35
Memory for Commissions. Test 1*.....	100	98	2
“ “ “ Test 2*.....	100	73	27
“ “ “ Test 3*.....	52	22	30
“ “ “ Test 3*.....	14	2	12
Weighted score on all four tests.....	(Max. diag. value)		35
Puzzle Tests.			
Healy Construction Puzzle A.....	“	“	“ 18
“ “ “ B.....	“	“	“ 35
Bicycle Bell Puzzle A*.....	“	“	“ 24
“ “ “ B*.....	“	“	“ 10
Puzzle A	“	“	“ 14
“ B*.....	“	“	“ 20
“ C*.....	“	“	“ 20
“ AC*.....	“	“	“ 27
“ BC*.....	“	“	“ 26
Learning from A to AC*.....	“	“	“ 11
“ “ B to BC*.....	“	“	“ 13
Pooled score on all nine puzzles.....	“	“	“ 34
Lifting the Table Asymmetrically Balanced*.....	“	“	“ 28
Suggestion by Progressive Lines			
Influence of Suggestion.....	“	“	“ 28
Accuracy in Reproducing Lengths.....	“	“	“ 19
Note.—Tests marked with an asterisk* are new.			

Note.—Tests marked with an asterisk* are new.

VII. CORRELATION OF ABILITIES WITH AGE

The essential feature of all quantitative measuring scales of intelligence is that they relate the total score of the individual to his age. The measure of the individual's intelligence is the relation between his performance and the average performance of other children of his own age. The various Binet scales and revisions compute the total score of the individual in terms of his "mental age." The difference between the "mental age" and the chronological age is used as a quantitative measure of intelligence. A variant of this measure is that obtained by dividing the "mental age" by the chronological age, the resulting "mental quotient" or "intelligence quotient" serving as a quantitative measure—a method advocated by Stern (62) and used very largely by Terman (65). Still another quantitative measure of intelligence is that used by Yerkes and his co-workers (82), in which the total score of the individual in a group of tests is referred to the averages of the scores of groups of similar individuals of different ages. The score of the individual compared to that of other similar individuals gives the "mental age" which compared with the chronological age will give the "mental status" (age difference) or the "coefficient of intellectual ability" (mental quotient) as quantitative measures.

If intelligence is measured in terms of age, the correlation of the tests with age should throw light on the correlation of the tests with intelligence. When we say that tests are diagnostic of intelligence, do we mean that they are diagnostic of age? Are those tests that show the most rapid growth with age those that have the highest value in differentiating groups of different intelligence? From the results of this investigation and that of Chotzen it is seen that different tests have different diagnostic values. The comparison of these results with results showing the growth of the abilities with age should throw light on the problems mentioned.

Results showing the correlation of the test abilities with age

are influenced by the error due to incomplete data. As far as the writer knows, Yerkes' results are the only ones that are free from this influence, for under the conditions of the application of the point scale, all of the tests must be given to all of the subjects. The present writer has therefore incorporated an analysis of Yerkes' data bearing on the problem of the relation between the individual tests and age.

In tables 30 and 32, (pages 123 and 125) Yerkes gives the results of 468 English speaking boys and girls from 4 to 15. The run of the 449 subjects from 5 to 14 is as follows:

Chronological ages	5	6	7	8	9	10	11	12	13	14
No. of subjects	28	55	48	47	43	53	55	40	43	37

The present writer has combined these data into five groups, 5 and 6, 7 and 8, 9 and 10, 11 and 12, and 13 and 14. In table 17 the results of each group are given on each test, the results being expressed in the form of the per cent. that the average number of points scored by each group is of the total number of points possible.

TABLE 17.
Growth of Abilities with Ages.
Per Cent. of Points Scored by Children of Various Ages.

Test.	5&6	7&8	9&10	11&12	13&14
1. Repeating sentences	64	65	67	74	71
2. Describing pictures	53	63	72	79	84
3. Repeating digits	55	65	73	80	84
4. Comparing lines and weights.....	51	67	86	97	100
5. Copying diamond and square.....	28	51	76	82	93
6. Defining concrete terms.....	36	47	65	74	78
7. Aesthetic comparison	68	87	99	100	100
8. Indicating omissions in pictures.....	57	75	88	97	98
9. Naming words	14	26	52	72	83
10. Comparing remembered objects.....	19	48	80	92	95
11. Counting backwards	14	59	97	98	99
12. Comprehending difficult questions....	11	24	43	60	80
13. Using three words in a sentence.....	1	8	57	77	84
14. Arranging five weights.....	21	53	79	92	89
15. Detecting absurdities	1	12	33	50	69
16. Line suggestion	31	46	64	76	83
16a. Length of letters.....	0	9	30	49	64
17. Defining abstract terms.....	0	0	21	43	67
18. Analogies	2	9	25	38	52
19. Drawing designs from memory.....	5	13	32	54	63
20. Reconstructing dissected sentences....	0	1	29	45	68

An examination of table 17 shows that the tests differ greatly in the rate of growth with age. The growth of the ability to repeat sentences, for instance, is represented by 64%, 65%, 67%, 74% and 71%, while at the other extreme are tests such as the test of counting backwards in which the growth is 14%, 59%, 97%, 98% and 99%. Further study of table 17 shows a high degree of similarity in the results of some tests. Take for instance tests 10 and 14

10. Comparing remembered objects	19%	48%	80%	92%	95%
14. Arranging five weights.....	21%	53%	79%	92%	89%

or tests 6 and 16

6. Defining concrete terms.....	36%	47%	65%	74%	78%
16. Resisting suggestion	31%	46%	64%	76%	83%

or tests 1, 2 and 3.

The similarities in the growth of the various tests are more easily shown graphically. In Fig 5 the writer has drawn the graphs of the various tests, having classified them roughly according to the similarities shown. All of the percentages were taken from table 32 (page 125), except that for test 2 age 5 which is obviously a misprint (38.7 instead of 48.7). The wide variation in the growth of abilities with age is clearly shown in Fig. 5. Tests 10, 11 and 14 show a very rapid growth, which is in marked contrast with that shown by tests 1, 2 and 3. Tests 9 and 13 show very nearly as rapid a growth as tests 10, 11 and 14, but the growth occurs for the most part between 8 and 11 or somewhat later than the abilities in the first group which are almost completely developed at 9. Test 5 shows a slope somewhat more gradual than tests 9 and 13, but slightly sharper than tests 4, 7 and 8 which are extremely easy for younger subjects. Tests 12, 15, 16a, 17, 18, 19 and 20 show considerable similarity. Test 12 is the easiest of the group and test 18 the hardest. Test 16a shows considerable variation, and as this was an extra test, it is possible that it was not given all the possible number of times. The slope of the curves of these seven tests is similar to that of tests 4, 7 and 8, but the abilities develop very much later, being hardly better than 25% at 8, having the fastest growth between 11 and 12 and scarcely reaching 75% in any test. The slope

of the curve for test 18 is more gradual than that of the other six tests, and probably coincides more closely with that of tests 6 and 16 which in turn have a more rapid growth than tests 1, 2 and 3.

Test 1 consisted of three sentences, one of 5 words (part a), one of 10 words (part b), and one of 18 words (part c), two points being awarded for the correct repetition of each. The explanation of the fact that test 1 shows practically no growth is given by Yerkes. "According to the results of our analysis of data, this test is eminently unsatisfactory, because parts (a) and (b) are so easy that even the four- or five-year-old child has little difficulty with them, whereas part (c) is so very difficult that only a few of the children among the 750 examined obtained credit for it. Such being the case, it is obvious that the score for this test cannot increase either markedly or regularly with increasing age" (page 128). In revising the scale, Yerkes changed the third sentence to 20 words, and added another of 15 words, giving credit of 1 point for repeating 5 or 10 words, and 2 points for repeating 15 or 20 words. With this system the growth of the test with age should be more rapid.

Yerkes finds nothing wrong with test 3. "Test 3 (Memory span for digits) has proved eminently satisfactory, and we see no reason for making other change than in position" (page 129). Comparison of the curves for tests 1 and 3 in Fig. 5 shows that test 3 is hardly more satisfactory than test 1. Test 3 consists of five parts, repeating 3, 4, 5, 6 and 7 digits, one point being allowed for the successful repetition of each. The lack of growth is evidently due to the same cause as test 1. The smaller numbers of digits are too easy, and there is little opportunity to differentiate superior ability. In curve A₃ the writer has inserted a graph of the average of the percentages given for non-selected Princeton boys and girls shown in table 13 of the first study under the discussion of sex differences (page 73). This test was weighted on a scale of six points, three points being awarded for repeating 7 digits, two points for 6 digits and one point for 5 digits. The advantage of this method of weighting tests according to their relative difficulty is obvious from comparing

curves A3 and 3 in plate IX. One method gives a differential measure of growth, the other does not.

It is fair to reason from a demonstrable error in tests 1 and 3 to a similar error in test 2. The three Binet pictures are shown in test 2, 1 point being allowed for enumeration, 2 points for description and 3 points for interpretation for each picture. The procedure in giving the test ("Look at this picture and tell me about it") was designed to avoid the response by enumeration suggested by the word "what," so that the diagnostic value of the test should in all probability be higher than that in the Trenton investigation. The lack of growth in the test is probably due to the fact that the gradations of response are not weighted according to their relative difficulty. Any child able to talk is able to enumerate, and this response hardly deserves credit. The present writer weighted the digits tests 1, 2 and 3 points because Goddard placed them in years VIII, X and XII. On the same line of reasoning, the three responses to the picture test would have the relative weight of 1, 3 and 7 rather than 1, 2 and 3, for they appear in III, VII and XV. This method is however entirely arbitrary, and it is a very simple matter to weight tests empirically according to their difficulty as shown in the discussion of the commissions test in the preceding chapter. If the three parts were weighted on a scale of 9 according to their relative difficulty as found in the Trenton investigation, the value for enumeration would be 0, for description 2 and for interpretation 7. With Yerkes' procedure in which enumeration is not suggested and description therefore easier, the relative weight of description and interpretation would be nearer 1 to 4 than 2 to 3.

The definitions test was given to 66 non-selected boys and girls age 6 to 7, and 50 non-selected boys and girls age 9 and 10 in Princeton, and to the normal and retarded 12, 13 and 14 year subjects in Trenton, the procedure and method of scoring being the same. The test of defining in terms of use was passed by 89% of the 6 and 7 year subjects, 96% of the 9 and 10 year subjects, and by all of the Trenton retarded and normal subjects. The test of defining in terms superior to use was

passed by 21% of the first group, 34% of the second group, and by 84% of the Trenton normal group, the performance of the retarded group (33%) being about the same as that of non-selected 9 and 10 year subjects (34%). Curves drawn for these three points are shown under A1 and A2 in Fig. 5. The test of defining in terms of use is probably as easy as test 7, while the test of defining in terms superior to use would seem to approach test 12 in difficulty. Curve 6 might be the resultant of two tendencies such as illustrated in curves A1 and A2. Yerkes gives 1 point for definitions by use and 2 points for definitions superior to use, 4 words being used. If the tests were weighed according to their difficulty, the proportion would be nearer 1 to 8 than 1 to 2. The effect of improper weighing is to obscure a real correlation with age.

The relations indicated by the growth of the other abilities are fairly definite. The test which shows the most rapid growth (test 11, counting backwards) was shown in the first section to depend upon school training. The curve of this test is not very different from the curve of tests 11 and 14 (comparing remembered objects and arranging five weights). It would seem that these tests were very valuable for indicating growth from 5 to 9, but as the abilities are practically developed at 9, they are useless as differential measures above this age. In the present investigation it was found that the five weight test was worthless for differentiating normal from retarded subjects, but there was reason to suppose that the test given to younger subjects would be diagnostic.

The test of constructing a sentence from three given words (test 13) shows a very rapid rise between 7 and 10 (60%) and a very much slower growth from 10 to 14 (23%). Yerkes gives credit of 2 points for a compound sentence and 4 points for a simple or complex sentence. His results indicate that very few children below 9 can construct a sentence at all. The part of the test that reveals the real growth is the mere construction of any sort of a sentence, the relative merits of the sentences constructed showing much less differentiation. This test needs checking up, for in the Princeton study there was

reason to suspect that the ability to construct any sort of a sentence possibly depended on the training of the third grade. The dependence of a test on school training may possibly be indicated by the rapidity of the growth of a test with age. Test 11 certainly shows a very rapid growth. Test 20 (dissected sentences) which depends partly on school training shows no growth till 9, a sharp rise at 9, but above this age agrees closely with other tests. This criterion of suddenness of growth is however very uncertain, as many other tests that have been shown to be independent of school training also show a very rapid growth. The proof of the presence of school training can only be shown by an analysis similar to that made in the first study.

It is probably not possible to reason from the rapidity of growth to the relative diagnostic value of the tests in a definite way. The curve of test 16, (the line suggestion test) which was shown to have a low diagnostic value is somewhat flatter than the curves of tests 12, 15, 20 etc., but this is an indefinite criterion. As a matter of fact the curve for test 16 shows two phases. There is a growth of 24% from 5 to 9, a growth of 26% from 9 to 10, and a maximum growth of 10% above 10. Yerkes gives credit of one point for each resistance, defining a "resistance" as saying "the same", or "equal", or for pointing to the left instead of to the right in the case of each of the last three pairs. In the present investigation this latter type of response was taken to indicate the influence of suggestion, a difference of procedure which probably accounts for the fact that the Trenton normal group resisted but 33% of the possible suggestions, while Yerkes' 12, 13 and 14 year subjects resisted 81%. Two characteristic responses to this test were pointed out (see page 56), the suggestion error and the discrimination error. It may be possible that subjects below 10 fall into the suggestion error, while those above 10 fall into the discrimination error. The character of the curve would indicate a change in the character of the response, or a change in the procedure in giving the test.

The comparison of the curve for test 18 (analogies) with that of test 12 (Comprehension) would show that the former was less useful than the latter in differentiating the intellectual growth from 5 to 14. Inferences from the slope of the curve to the diagnostic value are however uncertain. The design test was found to have a lower diagnostic value (36%) than the comprehension test (71%), dissected sentence test (71%), or the absurdities test (51%), yet the character of the curves for tests 12, 15, 19 and 20 are very much the same. The diagnostic value of the tests necessitates other evidence than that from the growth of the abilities, in the same manner as the demonstration of the presence of school training needs other evidence. However, it is perhaps possible to obtain corroboration of the diagnostic values found by comparing the performances of children of different ages.

It was noticed that the performance of the Trenton sub-normal group on the test of defining in terms superior to use was comparable to the performance of the 9 and 10 year non-selected Princeton subjects. Inasmuch as intellectual defect is usually regarded as a slowing up of mental development, the comparison of non-selected subjects age 9 and 10 with non-selected subjects age 13 and 14 might possibly throw some light on the relations found by comparing 12, 13 and 14 normal and retarded subjects. In the latter case the question was—what tests differentiate normal subjects of 12, 13 and 14 from retarded subjects of the same age? In this case the question is—what tests best differentiate children of 13 and 14 from children of 9 and 10? The comparison may readily be made by subtracting the third from the fifth column of table 17. The list of tests in the order of their value in differentiating 13 and 14 year subjects from 9 and 10 year subjects is as follows:

- 46% No. 17. Defining abstract terms.
- 39% No. 20. Reconstructing dissected sentences.
- 37% No. 12. Comprehending difficult questions.
- 36% No. 15. Detecting absurdities.
- 34% No. 16a. Length of letters.

31%	No. 9.	Naming words.
31%	No. 19.	Drawing designs from memory.
27%	No. 13.	Using three words in a sentence.
27%	No. 18.	Analogies.
19%	No. 16.	Line suggestion.
17%	No. 5.	Copying diamond and square.
15%	No. 10.	Comparing remembered objects.
14%	No. 4.	Comparing lines and weights.
13%	No. 6.	Defining concrete terms.
12%	No. 2.	Describing pictures.
11%	No. 3.	Repeating digits.
10%	No. 14.	Arranging five weights.
10%	No. 8.	Indicating omissions in pictures.
4%	No. 1.	Repeating sentences.
2%	No. 11.	Counting backwards.
1%	No. 7.	Aesthetic comparison.

For 13 of these tests (nos. 1, 2, 3, 6, 9, 12, 13, 14, 15, 16, 17, 19 and 20) values were obtained indicating their relative merit in diagnosing differences between normal and retarded subjects (shown in table 10.). The correlation (Pearson products-moments method) between the diagnostic value of these 13 tests found in this investigation and the value of these tests in differentiating 9 and 10 year subjects from 13 and 14 year subjects shown by Yerkes' results is 0.71 ($p = 0.09$). Four of the tests (nos. 1, 2, 3 and 6) show an error in the method of scoring. The correlation between Yerkes' results and the Trenton results for the other 9 tests is 0.81 ($p = 0.04$). The correlations are probably high enough to indicate that as a general rule the tests that most successfully differentiate normal subjects of 12, 13 and 14 from retarded subjects of the same age, also show the largest differences between the performances of 9 and 10 year subjects and 13 and 14 year subjects.

Concerning the results of the younger subjects, a similar question may be asked,—what tests most effectively differentiate 5 and 6 year subjects from 8 and 9 year subjects? The list of tests in the order of their value in making this differentiation is as follows:

83%	No. 11.	Counting backwards.
61%	No. 10.	Comparing remembered objects.

58%	No. 14.	Arranging five weights.
56%	No. 13.	Using three words in a sentence.
48%	No. 5.	Copying diamond and square.
38%	No. 9.	Naming words.
35%	No. 4.	Comparing lines and weights.
33%	No. 16.	Line suggestion.
32%	No. 15.	Detecting absurdities.
32%	No. 12.	Comprehending difficult questions.
31%	No. 7.	Aesthetic comparison.
31%	No. 8.	Indicating omissions in pictures.
30%	No. 16a.	Length of letters.
29%	No. 6.	Defining concrete terms.
29%	No. 20.	Reconstructing dissected sentences.
27%	No. 19.	Drawing designs from memory.
23%	No. 18.	Analogies.
21%	No. 17.	Defining abstract terms.
19%	No. 2.	Describing pictures.
18%	No. 3.	Repeating digits.
3%	No. 1.	Repeating sentences.

It is possible to compare these results on the easier tests with those of Chotzen. Chotzen (18) found that the backwardness of his feeble-minded children was most marked on nine tests, making change, naming months, recalling a story read, repeating sentences, repeating digits, defining in terms superior to use, counting backwards, comparing remembered objects and arranging five weights. The first three of these are not included in Yerkes' scale. There is an error in the scoring of the second three so that their value as a measure of growth is obscured. The last three show the highest value of any of Yerkes' 21 tests in differentiating 5 and 6 year subjects from 9 and 10 year subjects.

Chotzen also named nine tests as showing the greatest value in differentiating the groups of feeble-mindedness. These tests were naming coins, recalling a story read, making change, comprehending easy problem questions, repeating digits, defining in terms superior to use, copying a diamond, comparing remembered objects and arranging five weights. The first four of these tests are not in the point scale, and the value of the next two is obscured by the scoring system. The last three are included

in the five tests in Yerkes' scale that show the highest differential value for young subjects. The only test in the point scale showing a differential value for young children higher than 40% that is not included in Chotzen's list is that of constructing a sentence from three given words, and this test was not given a sufficiently large number of times by Chotzen. It is possible to conclude then that as a general rule the tests that most successfully diagnose mental defect in younger subjects, or most effectively differentiate the lower grades of feeble-mindedness, also show the greatest differentiation between the performances of 5 and 6 year subjects and 9 and 10 year subjects. The correspondence found between the results of the two studies of the diagnostic value of the tests and the results of comparing older and younger subjects are in agreement with the view that feeble-mindedness is a general slowing up of mental growth.

Again it is found that many of the tests that show the highest value in differentiating the higher grades of intelligence are tests that involve the use of language to a considerable extent. Yerkes' results afford the opportunity of studying the influence of language training on the tests. In table 31 (page 124) Yerkes gives the results of 196 children from 5 to 14 of non-English speaking parents. The present writer has computed the per cent. that the average number of points scored by the age groups 5 and 6, 7 and 8, 9 and 10, 11 and 12, and 13 and 14 was of the number of points possible to be scored. Subtracting these values from those given for similar groups of children of English speaking parents (shown in table 17) gives the influence of language training on each test in each group. These values are shown in table 18 in which the tests are arranged approximately in the order of the magnitude of the differences which they show between groups of different language training. A plus value indicates that the children of English speaking parents are ahead, a minus value that the children of non-English speaking parents are ahead.

The differences shown in table 18 vary from + 29% to - 9%, the median being + 4% ($Q=4.5\%$). There is then a general superiority of the results of children of English speaking parents.

TABLE 18.
Influence of Language Training.
Percentage Differences in Performance of Children of English and
Non-English Speaking Parents.

Test.	5&6	7&8	9&10	11&12	13&14
17. Defining abstract terms.....	0	0	+12	+18	+22
13. Using three words in a sentence.....	+1	+8	+22	+12	+8
14. Arranging five weights.....	-2	+29	+15	+4	+5
15. Detecting absurdities	0	+10	+14	+13	+13
12. Comprehending difficult questions.....	+6	+8	+7	+10	+19
20. Reconstructing dissected sentences.....	0	+1	+16	+8	+21
10. Comparing remembered objects.....	+4	+14	+11	+5	+2
18. Analogies	+1	+7	+10	+7	+10
9. Naming words	-1	+2	+5	+15	+13
6. Defining concrete terms.....	+3	+8	+9	+8	+5
16a. Length of letters.....	0	+9	+8	0	+15
1. Repeating sentences	+9	+8	+3	+6	-1
11. Counting backwards	-2	+10	+6	+2	+2
8. Indicating omissions in pictures.....	+11	+6	-1	+1	0
2. Describing pictures	+5	+1	+4	+1	+6
4. Comparing lines and weights.....	-5	+5	+10	+4	+2
3. Repeating digits	-5	-1	+4	0	+5
7. Aesthetic comparison	0	-3	+2	0	+2
5. Copying diamond and square.....	-6	-5	+3	-3	+2
16. Line suggestion	-4	-6	-9	+5	+2
19. Drawing designs from memory.....	-1	-1	-9	-3	+1

The correspondences shown in table 18 are of course more remarkable than the divergencies, but in the light of the high degree of correspondence, the fact of wide divergence would seem to indicate training. There are 19 differences higher than +10%, these differences being confined to but 10 tests. In three tests (nos. 8, 12 and 16a) differences higher than +10% occur in but one age group. In five of the tests (nos. 9, 10, 13, 14 and 20) these differences appear in two age groups. In two of the tests (nos. 15 and 17) these differences appear in three age groups. In five tests (nos. 12, 13, 14, 15 and 17) the average difference is +10% or higher. Although the evidence is thus directed against a few tests, it is probably impossible to say definitely on what tests the influence of language training is not present or on what tests the results are due to chance. The tests are arranged in table 18 in the approximate order of the magnitude of the differences found.

Two interpretations of the results are possible, the first, that the children of non-English speaking parents are under a serious handicap in some of the tests owing to deficient language training, the second, that these children have an inferior hereditary endowment, that they are less intelligent. The validity of the second interpretation may be examined by comparing the magnitude of the differences found between groups of younger and older children of English and non-English speaking parents with the magnitude of the values found for the tests in differentiating 13 and 14 year children from 9 and 10 year children, and for differentiating the latter group from 5 and 6 year children, or in other words by comparing the value of the tests as differential measures of growth with the supposed influence of language training.

As a measure of the amount of the influence of language training on younger subjects the differences found between the English and non-English speaking groups of 5 and 6 year children and 7 and 8 year children on each test were combined. The correlation (Spearman foot-rule) between these values and the magnitude of the differences between 5 and 6 year subjects and 9 and 10 year subjects is 0.11 ($p = 0.09$). As a measure of the amount of influence of language training on older subjects the differences found for the 11 and 12 year subjects and 13 and 14 year subjects were combined. The correlation between these values and the magnitude of the differences between 9 and 10 year subjects and 13 and 14 year subjects is 0.56 ($p = 0.09$). These values represent the relation between the differential value of the tests and the magnitude of the language differences at the extremes. The correlation between the differences found between the English and non-English speaking subjects of 9 and 10 and the differences between English speaking subjects of 7 and 8, and 11 and 12 is 0.35 ($p = 0.09$). In drawing inferences from these correlations it should be remembered that most of the tests that are the best differential measures of older children are passed by so few of the younger children that there is little opportunity of differentiating these children, or little opportunity for a language difference to appear. On the other

hand, some of the tests that show no differentiation between the performance of 9 and 10 year subjects and 13 and 14 year subjects fail to show a difference because they are too easy, and for the same reason these tests would fail to differentiate the language groups. These factors would tend to obscure the correlation between intelligence differences and language differences in the younger years, and to magnify this correlation for older subjects.

It is not possible to demonstrate whether the differences are due to deficiency in language training or to deficiency in intelligence. The statement that the connection between language differences and intelligence differences becomes more intimate with increasing years, (based on the correlations 0.11, 0.35 and 0.56) is modified by the relations pointed out between these correlations and the difficulty of the tests, and after all this may only be another way of saying that the tests that most successfully differentiate the higher grades of mental defect or most successfully differentiate the growth of older subjects involve language training. The position that the differences are due to language training is therefore favored. If the results justify the position that children of non-English speaking parents have an inferior hereditary endowment, it would appear that this inferiority becomes more marked with increasing age. However it must be remembered that this inferiority is only being measured by comparatively few effective tests. The safest conclusion would probably admit the possibility of both factors.

The examination of the results of the five weight test should indicate what the nature of the conclusions concerning the individual tests should be. It is perhaps surprising to find the apparent influence of language training in the test of arranging five weights. This influence is most marked in the younger years in which there is the strongest reason to believe that the differences are due to language training rather than to intelligence. Binet considered this test important in differentiating morons from normals, and attached considerable importance to it because it presupposed no acquired knowledge and was absolutely independent of all instruction. The results of the present in-

vestigation show that it is useless in differentiating the higher grades of mental defect, but there is reason to suppose that the intellectual factors of comprehending a serial arrangement and making the logically necessary comparisons are correlated with intelligence. Chotzen's results shown the test valuable in differentiating the lower grades of mental defect, and Yerkes' results show a very rapid growth from 5 to 9. It is significant that the same range in which the test is most effective is the range in which the influence of language training appears. If this influence is actually language training, it would indicate that one of the most important factors in the test, and the one that probably gives it its strongest correlation with intelligence is simply that of understanding the instructions. The results of the two language groups would indicate that the non-English speaking subjects, even though they are able to make the sensory discrimination and to comprehend and execute a serial arrangement, fail the test because they do not understand the instructions.

If the conclusion that the test of arranging five weights is influenced by language training is not justified, then it follows that none of the tests are influenced by language training, for the five weight test shows as marked an influence of this factor as any other test. The fact that the apparent language differences on this test are not due to differences in intelligence is indicated by the low correlation between the language differences and growth differences in this region (0.11), and by the fact that two of the tests that are most effective in making a differentiation of growth in this region (nos. 5 and 11) show very slight language differences. It must be concluded then that the differences are either due to language training or to chance. If the differences on the five weight test are due to chance, all other differences are due to chance. The chance hypothesis would probably be overworked in accounting for the fact that 76% of the 105 differences found between English and non-English speaking subjects were less than 10%.

If the foregoing analysis of Yerkes' data is correct it follows that some of the tests are influenced to a considerable extent

by language training. In the solution of some tests the children of non-English speaking parents are under a serious handicap owing to deficient language training. In any event the factor can not be disregarded entirely, for if there is some truth in the hypothesis that the language differences are due to intelligence differences, and that the difference in intelligence would manifest itself in the long run, in an individual examination how is the experimenter to know whether the subject's failure is due to defective intelligence or defective training?

In the present investigation the writer by selecting groups of similar language training was able to keep the data bearing on the diagnostic value of the tests free from the influence of the language factor. The position that the Trenton data were free from this influence is strengthened by Yerkes' results, for it was found in the present investigation that if it were to be concluded that certain tests depended on language training, it would also be necessary to conclude that two of the tests (naming 60 words and using three words in a sentence) showed this influence in favor of the children of non-English speaking parents, or in other words that the training in two languages was a positive help in these two tests. The fact that Yerkes' results show the English speaking children ahead in these tests would indicate that the language differences found in the Trenton study were due to chance rather than to the positive influence of the language factor.

VIII. RESULTS OF OTHER INVESTIGATORS

It is beyond the scope of the present investigation to summarize all of the literature bearing on the correlation of various mental tests with intelligence. The results of other investigators bearing directly on the individual tests used have been mentioned in the detailed discussion of the tests. Simpson's investigation is reported here because the problem has many points in common with that of this investigation and the conclusions have a similar trend. The other investigations of Norsworthy, Terman (earlier study), Wallin, and Pyle are mentioned only on account of the correspondence in the method of group differentiation. Workers in the field will find a summary of the work bearing on the correlation between many of the standard tests and intelligence given under the heading "Dependence on intelligence" under each of the tests in Whipple's Manual (76). An analysis of the factors involved in many of the better known tests and the relation of these factors to intelligence is given in the second volume of Meumann's "Vorlesungen" (45), along with an evaluation of the results of various investigators who have applied the tests in this particular field. Lastly, for the real masterpieces in the creative portion of the field, the reader is referred to Binet's original articles which appeared from time to time in *L'Année Psychologique*, and which have recently been translated into English by Kite (39 and 40).

Simpson (58) gave 15 tests to two groups of adults who were taken to represent "the two extremes of 'general intelligence' as judged by the world", one group being composed of 17 professors and advanced students, and the other of 20 men who had never held any position demanding a high grade of intelligence. The tests used included two tests of perception (marking A's and marking geometrical forms), three tests of memory (memory of unrelated words, of passages, and recognition of forms previously seen), four of association (addition, easy opposites, learning pairs of words and forms, and com-

pleting words), three of selective thinking (hard opposites, mutilated prose, and absurdities), two of sensory discrimination (reproducing lengths and discriminating lengths), and one of motor control (scroll test).

Simpson found that "the tests reveal very marked differences in the two groups in language tests demanding selective thinking; marked but less difference in certain tests of memory; very decided differences in language tests demanding speed and accuracy in easy association; less difference in the more directly practiced and mechanical associations demanded in adding; in perception tests and in motor control the differences are somewhat less still; and in discrimination of lengths they are least of all." (page 55).

The method on which Simpson based his conclusions was that of comparing the per cent. of the poor group that surpassed the median of the good group, and the lowest four, two and one of the good group. From the results given in table II, (pages 30 and 33 of Simpson's monograph) the present writer has calculated the Maximum Diagnostic Value of each test. The list of tests in the order of the Maximum Diagnostic Value is as follows:

- 100% Test IV. Easy Opposites.
- 100% Test XII. Hard opposites.
- 94% Test XIV. Ebbinghaus' mutilated text.
- 94% Test V. Recognizing forms.
- 88% Test VII. Learning pairs.
- 85% Test VI. Memory for words.
- 78% Test VIII. Memory for passages.
- 75% Test XIII. Completing words.
- 73% Test III. Scroll test.
- 70% Test XI. Adding.
- 70% Test I. Marking A's.
- 69% Test II. Marking geometrical forms.
- 42% Test X. Estimating lengths.
- 21% Test IX. Reproducing lengths.

The results of the absurdities tests were not used because they were not reliable.

In drawing conclusions concerning what sort of abilities were

connected with mental ability, Simpson used the intercorrelations of the tests as a basis of classifying them. The Ebbinghaus and hard opposites tests, for example, were classified under "selective thinking", and the easy opposites test under "association". The correlation between the first two tests was 85, so that they may be correctly classified under one heading. The easy opposites test correlated higher than these two tests of "selective thinking" (72 and 83) than with any of the other tests, so that it should properly be placed under this heading rather than under "association".

Grouping the tests together on a basis of their intercorrelations, Simpson figured the average correlation of each ability with all of the other tests. On the basis of the magnitude of these average correlations, Simpson concluded that "power of selective thinking is more intimately connected with, and more characteristic of, general mental ability than is any of the other abilities tested; that memory is next most highly correlated with general ability; the simpler forms of association next; perception next; motor control considerable less; and discrimination of lengths least of all." (page 67).

Simpson held that the tests were measures of mental capacity rather than measures of amount of training and education, because the correlation between the number of years of schooling and the rank in the eight tests that correlated most highly with the other tests was low (38). On the evidence of studies of retardation he held that "a small number of years schooling means inability to learn advanced and difficult language work, rather than lack of opportunity to learn it." (page 70). Further evidence that the tests of selective thinking were not measures of school training was derived from the fact that the subjects who were considered decidedly dull or stupid by their fellows did poorest in these tests. Simpson found further evidence in support of the proposition that language tests are fair tests of ability from the fact that the intelligence of primitive peoples may be measured by their language, and that feeble-minded children are deficient in acquiring higher forms of language.

A combined measure of general intelligence was taken for

each individual by adding his scores on the Ebbinghaus test, hard opposites, easy opposites, learning pairs and recognizing forms, the scores being compared in terms of the deviation from the median. The deviations of the good group varied from + 95 to + 21, while those of the poor group varied from 0 to - 127. There was a difference of 21 between the lowest member of the good group and the highest member of the poor group, this difference being 46.5% of the average deviation from the median (45.14). The combined score therefore differentiated the groups more completely than the score on any of the individual tests. Further evidence that the combined score on these five tests was a measure of "general intelligence" was obtained from the fact that the correlation between the ranking of the subjects of the good group according to these tests and their ranking according to the independent estimates of their intelligence by ten or more persons was 92. The correlation between the various tests and estimated intelligence varied from 96 (hard opposites) to - 20 (drawing lengths).

Norsworthy (48) gave twelve mental tests and four physical tests to 150 defective children and to large numbers of normal children in order to determine whether the mental defects of idiots were equalled by the bodily defects, whether idiots formed a separate species or not, and whether idiots showed a lack of mental capacity all around. The physical tests showed very slight differences between the groups, 26% of the idiots being above the median of the normals in the measurements of temperature, 44% in weight, 45% in height and 49% in the measurements of pulse. The median for the idiots in "intelligence tests" (part-whole, genus-species and opposites tests) was below the median of the normals 7 times the probable error. The median for idiots in memory tests (memory for dictated passages and related words) was below the median of normals 3.5 times the probable error. The median of idiots in "maturity tests" (reproducing a weight to a standard, memory for unrelated words, cancelling A's and cancelling a's and t's) was below the median of normals 2.7 times the probable error.

Norsworthy finds no evidence for the theory that idiots con-

stitute a separate species of individuals. The results show however that the differences that are found between idiots and normal children vary with the measurements used, which is of course another way of saying that the tests vary in their efficiency in diagnosing feeble-mindedness. Physical tests have little efficiency, and the controlled association tests (combined score) had twice the efficiency of the two memory tests.

Terman (64) examined seven of the brightest and seven of the dullest pupils in a school system made up of about 500 children, finding the bright boys superior in all the mental tests given, but below in the motor tests. No suggestions can be obtained concerning the relative efficiency of the tests in differentiating the groups on account of the small number of subjects.

Much information concerning the problem of what tests are diagnostic of intelligence may be gained from the correlation methods in which the standing of a group of subjects in a series of tests is compared with their rank order in intelligence as estimated by the school-masters, school-fellows, or other persons supposedly competent to diagnose mentality. The results of this method are of course no more accurate than the original independent rating of intelligence, and this rating is not absolutely reliable for the correlations between the ratings of one group by different observers are frequently low. Furthermore, there is danger that the individuals who make the ratings will stress some ability such as memory so that some tests will show a correlation with estimated intelligence somewhat higher than their probable true correlation with intelligence.¹

On the whole, these correlation methods are chiefly serviceable in determining the relationships between the various test abilities. Eventually, if the same tests are given to different groups,

¹ This point is well illustrated by Abelson's (1) results. Abelson instructed the teachers to estimate the "practical intelligence" of the children by considering in forming their opinion which of the children they would soonest trust on an errand requiring the sharpest intellect. The fact that the test that showed the highest correlation with intelligence was a test of memory for commissions shows that the teachers considered the mere retention of the instructions more important than the intellectual factors involved in the execution of the errand.

the evidence from different investigations will be very valuable in showing what tests are most diagnostic of intelligence. At present there are not enough investigations available. The work to date however would certainly support the two propositions that different tests vary in their effectiveness in diagnosing intelligence, and that a combined score of several tests is more effective than any single test.

Wallin (73) gave the Binet 1908 scale to a large group of epileptics, and compared their results with those of other investigators on normal and feeble-minded individuals. Certain tests proved to be especially difficult for epileptics just as certain tests in this investigation proved to be especially difficult for the retarded group. The results of the two investigations cannot be compared however, for Wallin referred the discrepancies between the performance of normals and epileptics to "inherent abnormalities in the mentation of the epileptics", and indeed there is no reason for believing that the tests that proved especially difficult for epileptics should also be the tests that are most highly diagnostic of feeble-mindedness.

Pyle (53) gave a series of class-room tests to groups of pupils classified as bright and dull on the basis of their school marks. He found that the completion, word-building, logical memory and controlled association tests were most valuable for the purpose of ascertaining the mental differences between the groups. Ability to do the cancellation test in some cases showed an inverse relation to the other tests. Ability in the ink-blot test showed an inverse relation with age, (the younger children doing better), and showed a negative relation to the other tests.

Terman (65) has recently published a revision of the Binet scale, the selection of tests being based on an empirical verification of their validity. The method of demonstrating the validity of the individual tests was that of comparing each test with the scale as a whole. The subjects of each age were divided into three groups according to their "intelligence quotients" (IQ), and the tests that showed a higher per cent. passed in an inferior IQ group than in a superior IQ group were rejected. This method insures that each test is to some extent

coherent with the scale as a whole. The results of this method are best shown by the following quotation from Terman:

"When the tests were tried out in this way it was found that some of those which have been most criticized have in reality a high correlation with intelligence. Among those are naming the days of the week, giving the value of stamps, counting thirteen pennies, giving differences between president and king, finding rhymes, giving age, distinguishing right and left, and interpretation of pictures. Others having a high reliability are the vocabulary tests, arithmetical reasoning, giving differences, copying a diamond, giving date, repeating digits in reverse order, interpretation of fables, the dissected sentence test, naming sixty words, finding omissions in pictures, and recognizing absurdities." (Pages 76 and 77).

"Among the somewhat less satisfactory tests are the following: repeating digits (direct order), naming coins, distinguishing forenoon and afternoon, defining in terms of use, drawing designs from memory, and aesthetic comparison. Binet's "line suggestion" test correlated so little with intelligence that it had to be thrown out. The same was also true of two of the new tests which we had added to the series for try-outs." (Page 77).

"Tests showing a medium correlation with the scale as a whole include arranging weights, executing three commissions, naming colors, giving number of fingers, describing pictures, naming the months, making change, giving superior definitions, finding similarities, reading for memories, reversing hands of clock, defining abstract words, problems of fact, bow-knot, induction test, and comprehension questions." (Page 77).

From the standpoint of the desirability of comparing Terman's results on the individual tests with those of this investigation it is to be regretted that Terman's actual data are not yet available. However, he uses the agreement of each test with the scale as a whole as a criterion of the test's correlation with intelligence, and from his report it is possible to classify the tests into four grades of reliability, those showing a high correlation, a medium correlation, a less satisfactory correlation, and no correlation with intelligence, (the term "correlation with

intelligence" being used interchangeably by Terman with "correlation with the scale as a whole").

It is not possible to compare all the results of this investigation with those of Terman on account of the differences of procedure. The agreement between Terman's procedure and that of the present investigation was very close on nine tests, and a direct comparison is possible. Terman found that the dissected sentence test showed a high reliability, and this test showed one of the highest diagnostic values found in the present investigation (71%). He also found that the tests of naming 60 words, giving rhymes with "day", "mill" and "spring", and naming the date showed a high correlation with intelligence, while the diagnostic values found in this investigation were not particularly high (40%, 32% and 29% respectively). The tests of naming the months and arranging five weights which Terman reported as showing a medium correlation with intelligence have diagnostic values of 43% and 18% respectively.* Two tests that are classified as less satisfactory, the designs and 7 digits tests, show diagnostic values of 36% and 30% respectively. The line suggestion test which correlated so little with intelligence that it had to be thrown out showed one of the lowest diagnostic values found in this investigation (12%).

Five other tests used in this investigation may be compared with Terman's results although the procedure was somewhat different. Terman used five absurdity questions, three of them being the same as the present writer used. Terman reported this test as showing a high correlation with intelligence, and considered it "one of the most ingenious and serviceable tests in the scale", and "an invaluable test for the higher grades of mental deficiency", an opinion in keeping with the diagnostic value found in this study (53%). Terman used four grades of comprehension questions, and found that the test showed a medium correlation with intelligence. This test showed one of the highest diagnostic values in the present investigation (71%). The two studies used different words in the test of defining terms superior to use so that the results are but roughly com-

parable. Terman classified the test as having a medium correlation with intelligence, while the value found in this investigation was 51%. Two of the three words used in the abstract definitions test appear in the five used by Terman. The diagnostic value found was 51%, and Terman reported it as showing a medium correlation with intelligence. Terman used three problems from various facts, and found a medium correlation with intelligence. The two problems used in this study showed different diagnostic values (21% and 51%). It is not possible to compare Terman's results on the tests of describing and interpreting pictures with those of this investigation because the pictures used and the procedures in giving the tests were different.

The results of the two investigations, where comparison is possible, do not agree very closely. Tests that showed a high correlation with intelligence according to Terman showed diagnostic values of 71%, 53%, 40%, 32% and 29% in this study. The diagnostic values found for the tests classified as showing a medium correlation were 71%, 51%, 43% and 18%. The tests classified as less satisfactory showed values of 36% and 30%, while the test that was so unsatisfactory that it had to be eliminated showed a value of 12%.

The discrepancies between the results of the two investigations might possibly be explained by the difference of method. In this event the question arises as to which method is more reliable. A. S. Otis (49) points out that it is theoretically possible to have a coherent system of tests that are not tests of intelligence (tests of physical strength, for example), and that it is therefore necessary to have other criteria of the validity of the individual tests. This objection is of course largely theoretical, but it is possible that the tests in different portions of the scale vary in the degree in which they correlate with intelligence or depend on factors other than intelligence so that the criterion of coherency would not give results that were constant throughout the scale.

The 16 tests in years VIII and IX, for instance, include the tests of counting backwards from 20 to 0, naming coins, giving the date, making change, naming months and counting stamps

which were shown to depend on school training, the test of writing from dictation which Binet eliminated on account of school training, and the test of constructing a sentence from three given words, a test in which this factor was suspected. The large proportion of tests in this region (8 out of 16, or if alternatives are omitted 4 out of 12) that are dependent on training might account for Terman's finding that some of the tests "which have been most criticised have in reality a high correlation with intelligence." Logically the test of coherency would indicate dependence on training as much as on intelligence. In the light of the first study, the presumption would be that the test of coherency in the region of VIII and IX would show the tests of training to have an abnormally high validity. The training tests may be diagnostic of intelligence, but that is another matter. The objection against the test of coherency is that it fails to take account of variable factors.

In the absence of the actual data it is not possible to determine the cause of the discrepancies between Terman's results and those of this investigation. At least Terman may be considered as subscribing to the general thesis that the individual tests vary in the degree in which they correlate with intelligence or in their value in diagnosing intelligence.

IX. CONCLUSIONS AND SUGGESTIONS

In all twenty-three Binet tests were used in this investigation, it being possible to draw conclusions concerning nineteen of them. The diagnostic values of these nineteen tests and their various sub-tests are shown in table 10. Ten supplementary tests were used, it being possible to draw conclusions concerning nine of them. The diagnostic values of these nine tests and their various sub-divisions are shown in table 16. Table 19 shows the various Binet tests and the supplementary tests arranged in the order of their diagnostic value as shown by this investigation.

The extreme divergency of the tests is clearly shown in Table 19. Three tests show diagnostic values higher than 70%, four higher than 60%, eight higher than 50% and twelve higher than 40%. Twelve tests show diagnostic values lower than 30%, five less than 20% and two less than 10%. The writer does not insist that all the values given in table 19 are absolutely final and definite. Indeed the influence of the one variable factor in the results, the personal equation, is so subtle that it can hardly be avoided. The experiment has however been reported in detail so that it can be repeated.¹

The reader may draw his own conclusions concerning the nature of the tests that are most diagnostic of intelligence, or the nature of the mental processes most intimately connected with intelligence. Inferences from the nature of the tests to the nature of intelligence are of course uncertain, for we know very little about the mental processes involved in the tests. The mere fact that a psychologist classifies a test as involving a certain process does not prove that that process is involved. In a general way it is perhaps interesting to note that there is

¹ The writer will gladly communicate further details of procedure that have not been reported to anyone wishing to repeat the experiment. All the data from the Princeton and Trenton experiments are on file at the Princeton laboratory, and are available for anyone who wishes to check up the writer's computations or to make further calculations.

TABLE 19.

List of Tests in the order of their Diagnostic Value.

- 74 Subtraction tests.
- 71* Comprehending difficult questions.
- 71* Reconstructing dissected sentences.
- 64 Healy cross-line tests.
- 53* Detecting absurdities in statements.
- 51* Defining in terms superior to use.
- 51* Defining abstract terms.
- 51* Solving problems from various facts (Problem b).
- 49 Balance test.
- 49 (42) Distinguishing between terms.
- 43* Enumerating the months.
- 40* Naming 60 words in three minutes.
- 36* Copying designs from memory.
- 35 Estimating lengths.
- 35 (30) Memory for commissions.
- 34 Puzzle tests. (Pooled score.)
- 33* Giving rhymes with "defender."
- 30* Repeating 7 digits.
- 29* Using three words in a sentence (2 ideas).
- 29* Giving the day and date.
- 28 Lifting the table asymmetrically balanced.
- 28 Influence of suggestion. (Suggestion by progressive lines.)
- 23* Repeating a sentence of 18 syllables.
- 23* Using three words in a sentence (1 idea).
- 21* Solving problems from various facts (Problem a).
- 19 Reproducing lengths. (Suggestion by progressive lines.)
- 18* Arranging five weights.
- 12* Resisting suggestion.
- 6* Interpreting pictures.
- + 2* Describing pictures.

Note: Tests marked with an asterisk (*) are in the Binet series.

much in common between Stern's (62) definition of intelligence as "general mental adaptability to new problems and conditions of life", Witmer's (74) definition of intelligence as "the ability of the individual to solve what for him is a new problem", and Pillsbury's (52) definition of reasoning as "the

application of any knowledge in a new way". Intelligence tests would seem to involve a new problem and a solution of the problem, or in other words, reasoning.

One logical implication of this study should be pointed out. Throughout the study the emphasis has been placed on the diagnostic value of the tests, or their merit in differentiating feeble-minded individuals from normal individuals. If the conception of diagnostic tests is carried to the extreme, the belief that certain tests could be found that are absolutely diagnostic of feeble-mindedness would imply that feeble-minded individuals constituted a separate species or a group of individuals who were in some respects completely different from normal individuals. Of course it has never definitely been shown that feeble-minded individuals do not constitute a separate species in some respects. Norsworthy's (48) results negate this view to some extent but not conclusively, for her results show feeble-minded individuals to be more distinct in some respects than in others. Logically of course there are no degrees of being a species, but there are degrees of accuracy of definition by which a species is specified, and Norsworthy in concluding that feeble-minded individuals did not constitute a separate species drew these conclusions on the basis of the tests used.

The view that feeble-mindedness is a general slowing up of mental development gets its chief impetus from the convention started by Binet and followed by others of defining normal development in terms of age, and this view of the intelligence of the feeble-minded if true would completely disprove the view that they were a separate species, unless intelligence changes in character in the course of its development. Theories of the correlation of intelligence with age have all been based on cross-section studies of different individuals at different ages, and the true nature of the development of intelligence will probably not be known until longitudinal studies of the same individuals through a number of years have been made. The results of this investigation show that as a general rule the tests that were most effective in diagnosing known differences

of intelligence also showed the greatest differentiation between subjects of different ages. The correlation found was not absolutely valid but in general supported the view that feeble-mindedness is a slowing up of mental development. These findings do not absolutely support this view for there may be changes in the character of intelligence with increasing age. The results of this investigation neither establish nor controvert the view that feeble-minded individuals in some respects constitute a separate species, a view the validity of which the belief in the possibility of discovering tests that are absolutely diagnostic of feeble-mindedness would necessarily imply. In the absence of definite experimental results the discussion of this point is largely speculative.

As a practical matter it is significant that the results of this investigation show that many of the tests that are diagnostic of the higher grades of mental defect involve the use of language. Six of the tests that show diagnostic values higher than 40%, abstract definitions, absurdities, comprehension, dissected sentences, 60 words and concrete definitions, stand first, fourth, fifth, sixth, ninth and tenth respectively in the list of tests in Yerkes' point scale arranged in the order of the magnitude of the differences found between English and non-English speaking children. The other six tests showing diagnostic values higher than 40% are not in Yerkes' scale. Two of the remaining four tests in the first ten of Yerkes' list (five weights and sentence test) show no diagnostic value in this investigation, and the other two (comparison and analogies) were not used. The character of two of the other tests showing diagnostic values higher than 40% (solving problems and distinguishing between terms) would indicate that they might be influenced by language training. Many of Simpson's (58) tests that most effectively differentiated his groups would also seem to involve language training.

Many of the tests that most effectively differentiate the higher grades of mental defect involve the use of language. In view of the probable close connection between intelligence and reason-

ing ability, it is not surprising to find a close connection between reasoning and the common vehicle of its expression, language. Of course it is possible to have reasoning in action as well as in thought but the fact remains that we have very few tests of reasoning in action.

The dearth of intelligence tests that are independent of the language factor indicates the magnitude of the problem that faces American investigators who wish to test the intelligence of individuals who have not had adequate training in English, immigrants, children of non-English speaking parents etc. It is possible to find many tests for older persons that do not involve language but a great many of these tests do not involve intelligence. With younger children and lower grade cases, the influence of language may appear in the instructions rather than in the test itself, as in the case of the five weight test. The problem of testing individuals without adequate training in English has considerable practical importance in this country with its cosmopolitan population. The solution of the problem will most certainly involve many careful and skilful researches.

To one who has followed the analysis of the various intelligence tests through these pages, there is apparently a hopeless confusion in the field. Some tests involve the influence of the personal equation to a marked degree, others depend on school training, some depend on linguistic training, and far too many depend too little on intelligence. The only hopeful aspect of the situation is that it is possible to place the whole field of mental tests on an absolutely empirical basis.

The personal equation is a difficulty but not an insurmountable difficulty and the presence or absence of this factor may be empirically demonstrated. The influence of scholastic training may be determined by comparing groups of similar ages but different training. The influence of sex differences may be determined by comparing the results of boys and girls of the same age and with the same training. The dependence of the tests on language may be determined by comparing the results of groups of English and non-English speaking children with

the same environmental opportunity. And lastly, the reliability of a test as a measure of intelligence may be determined by giving it to groups of the same age and opportunity, but of known differences of intelligence. There is no room for a priori objections or inferences. Every factor may be empirically demonstrated. It is possible to construct scales for measuring intelligence without knowing exactly what intelligence is. A person may determine the presence or effectiveness of intelligence without knowing its nature.

In view of the fact that the merit of tests and systems of tests may be empirically demonstrated, it is legitimate to demand that the investigator who proposes a new scale or another revision of existing scales should offer a demonstration of the reliability of his method. If he has a scale for measuring the higher grades of defect, let him show that it will actually diagnose these conditions. If he has a scale for differentiating adolescents from adults, let him demonstrate that the scale will actually make this differentiation.

It is surprising that up to this time very few complete demonstrations have been made of the reliability of measuring scales of intelligence. Binet's only experimental verification of his scale consisted in showing that the distribution of the children testing "at age" and below and above age was normal.² As a matter of fact there was practically no experimental verification of the scale, and its validity rests on Binet's merit as a psychological observer. The fact that some of the tests are worthless merely proves that Binet was occasionally mistaken in his opinion, and the fact that so many of the tests are extremely valuable is a lasting tribute to his experimental genius. Goddard's statistical verification of the Binet scale has been shown to be faulty by Ayres (2), Schmitt (56), Thorndike (69), Yerkes (82), and others.

Yerkes evidently revised the Binet scale on a point basis without experimental verification of his opinions. He threw out some tests because he thought they depended on school train-

² The totals of the table demonstrating this fact do not add up correctly. *Année Psychologique*, 1908, Vol. 14, page 73.

ing, and weighted the tests and their various parts on an entirely arbitrary basis. Of course Yerkes guessed very shrewdly, for only two of the tests in the scale involve school training, as a general rule the most valuable tests receive the most weight, and only three or four of the tests are worthless on account of faulty weighting. So far as the writer knows, Terman (65) is the only investigator who has offered an empirical verification of the individual tests along with the publication of the system of tests. Terman's method of demonstrating the validity of the individual tests has been discussed in the previous chapter.

In affording an experimental verification of his scale the experimenter should in every case make his demonstration from the individual tests, and not from the total score of the whole system, for the only source of progress in the perfection of systems of tests lies in perfecting the individual tests. The system as it stands may be fairly effective, but the analysis of the individual tests will usually show that it could be more effective.

Every study of systems of tests will probably show that the total score has greater reliability than any of the individual parts. The twenty-three Binet tests used in this investigation scored together in the form of "mental ages" had a maximum diagnostic value of 69%, which is more than double the average diagnostic value of the individual tests. Yet if the total score were computed from 5 of the most effective tests, the diagnostic value was 83%. The same has been indicated time and again. The pooled score always shows a higher correlation with intelligence than the individual tests. The whole is more reliable than the parts, yet the effectiveness of the whole is raised by increasing the reliability of the parts.

To one who has followed the analysis of the Binet tests through these pages, it is perhaps surprising that the scale works. Yet it does work within certain limits. It will indicate pronounced defect in children over certain ages, and as the writer has shown in a previous article (14), it will diagnose the finer shades of intelligence from 7 to 11 as expressed by

the teachers' judgments. The merit of the scale as it stands undoubtedly rests on two principles. In the first place, the tests are arranged in the approximate order of their increasing difficulty (by the criterion of 75% passed) so that the experimenter can find tests within the subject's range of ability by exploring from the "basal age" upwards. In the second place, the experimenter in exploring this ability gives the subject a number of tests. The merit of the scale rests then on the principles of having a number of tests, and of having those tests within the range of the subject,—neither too far above or too far below his ability. There is nothing essentially new in these principles. It is a matter of common observation that the user of the shot-gun frequently chooses the "spread" instead of the "choke" barrel, and that the user of the rifle invariably adjusts his sights to suit the range.

In regard to the principle of having a number of tests, Binet is quite frank. His explanation of the reason for measuring intelligence by groups of tests follows. "Obviously it rests upon the principle that a particular test isolated from the rest is of little value, that it is open to errors of every sort, especially if it is rapid and is applied to school children; that which gives a demonstrative force is a group of tests, a collection which preserves the average physiognomy. This may seem to be a truth so trivial as to be scarcely worth the trouble of expressing it. On the contrary it is a profound truth, and good sense is so far from being sufficient to divine this so called triviality, that up to the present it has been constantly disregarded. One test signifies nothing, let us emphatically repeat, but five or six tests signify something. And that is so true that one might almost say, 'It matters very little what the tests are so long as they are numerous.'" (Kite's (39) translation pg. 329).

Inasmuch as no article on mental tests is complete without a suggested revision, the writer will indicate his opinion on the nature of such a revision. If the presence of so many variable factors in mental tests means that it is impossible to have quantitative measuring scales of intelligence, then of course

the development of mental tests is merely the development of more diagnostic tests with studies of the types of performance that are qualitatively diagnostic. However, the assumption underlying a qualitative diagnosis is that owing to certain facts the experimenter expects the subject to do differently. The certain facts that the experimenter takes into consideration represent merely the performance of other similar children. In other words the assumption underlying a qualitative diagnosis is the assumption of norms of performance, and if we can have norms of performance on individual tests, we can have norms on groups of tests with the probability that the group of tests will be more reliable than its components. The assumption that norms of performance are possible rests on the assumption of equality of opportunity. The following remarks are made on the assumption that it is possible to classify individuals as having had equal opportunity.

One of the first principles to be recognized in considering what the nature of an adequate measuring scale would be is that the individual tests should be scored on the point or partial credit system rather than on the all or none system, for as Yerkes points out, the scoring of tests by points brings out the full value of the testing material and minimizes the influence of the personal equation. It might be objected that some tests can not be scored on a basis of partial credits, for they are all or none tests, they are either passed or failed. The months test and the counting stamps tests are examples of all or none tests. It is probably true that no test should stand entirely by itself, as there is always a danger that the subject has been told about some individual tests. If the tests of naming the months, days of the week, giving the date etc. are diagnostic tests, their efficiency would not be decreased by combining them into one test and weighting them according to their relative difficulty, or if the ability in the counting stamps test is diagnostic, a person would stand a better chance of measuring this ability by giving other problems than that of counting three 2 cent stamps and three one cent stamps (counting two 2's and two 1's, three 3's and three 1's, for example). It has been shown

that there are dangers in weighting parts of tests according to the point system, but the parts may be weighted empirically according to their difficulty.

Another principle that should be recognized in considering the nature of an adequate measuring scale is that a number of tests should be given well within the range of ability of the subjects tested, for the best differential measure of a number of groups is one well within the ability of the groups. This was brought out in the studies of the personal equation, grade correlations, sex differences, and diagnostic value of the tests. It is an obvious principle, but one that needs recognition. An example from Yerkes' point scale will make this clear.

The examination of Fig. 5 in which the growth of each test in the point scale with age is shown graphically shows that for the most part the tests are either one thing or another—they are either tests for young children or for older children. Tests 4, 5, 7, 8, 10, 11 and 14 are useful for indicating growth from 5 to 9, but are practically useless beyond that point for the abilities are almost completely developed at 9. The 9 year subjects scored over 75% of the possible number of points on these seven tests, and there is less than 20% improvement manifested above 9. At the other extreme are tests 12, 15, 16a, 17, 18, 19 and 20 that are apparently valuable for indicating growth above 9, but are practically useless below 9. Of the remaining 7 tests, 1, 2, and 3 are useless for all ages, tests 6 and 16 have doubtful significance, and tests 9 and 13 alone have value for differentiating the intermediate growth from 8 to 12.

From the data in table 30, page 123, the present writer has calculated norms for each year for four point scales, (1) the original point scale without test 16a, (2) the same scale eliminating tests 1, 2, 3 and 6 that have an error in the scoring, (3) a scale for younger children consisting of tests 4, 5, 7, 8, 9, 10, 11, 13 and 14, and (4) a scale for older children consisting of tests 9, 12, 15, 16a, 17, 19 and 20. The norms are calculated on the basis of the per cent. that the number of points scored is of the number possible to score, that number being 100 points for the first scale, 72 points for the second scale, 34 for the third and 36 for the fourth. These norms are given in table 20.

TABLE 20.
Norms for Four Point Scales.

Chronological ages	5	6	7	8	9	10	11	12	13	14
Scale 1.....	22	30	36	42	56	62	65	77	79	81
Scale 2.....	12	20	27	35	51	59	62	75	79	82
Scale 3.....	22	35	45	57	76	82	85	93	93	94
Scale 4.....	3	6	10	14	31	38	43	63	70	73

The effect of throwing out tests 1, 2, 3 and 6 is to lower the norms of the younger children without changing those of the older children to any extent, so that the subjects may express their ability within a range of 70 points instead of 59 points. In other words the scale is more effective, for the greater the range in which differences may be expressed, the greater the possibility of differentiation. In the same way scales 3 and 4 are more effective than scale 2. Scale 2 gives a range of 39 points between 5 and 9, while scale 3 gives a range of 54 points. Scale 2 gives a range of 31 points from 9 to 14 while scale 4 gives a range of 42 points. Scale 1 has a range of 59 points, 58% of the increase being scored by children from 5 to 9 and 42% by children from 9 to 14. Scale 2 has a range of 70 points, 56% being scored by the younger group and 44% by the older group. On the other hand, 75% of the 72 points in scale 3 are scored by the younger group (5 to 9), and 60% of the 70 points in scale 4 are scored by the older group (9 to 14).

The foregoing demonstration does not mean that Yerkes' scale as it stands would be more accurate if broken up into two parts, for the accuracy is in some measure dependent on having a sufficiently large number of tests. It is true nevertheless that the use of tests above or below the ability of any group lessens the possibility of differentiating that group. The accuracy of any point scale system of testing intelligence increases with the number of parts, in the degree in which it tends away from universality. In a recent communication to the "Symposium" in the *Journal of Educational Psychology* (63) Yerkes stated that he had abandoned his plan for a universal point scale, and suggested three age scales, from birth to 4 years, from 4 to 12, and from 12 to maturity or 16. Ultimately, he will probably have to split the scale from 4 to 12 into two scales at least.

Aside from the fact that the difficult tests decrease the possibility of differentiating the younger children, and the easy tests decrease the possibility of differentiating the older children, it is a waste of the experimenter's time to ask a young child to perform tests obviously beyond his ability, or to ask an older child to do tests way beneath his ability. But the experimenter never knows the range of a child's ability till he starts to test. What is needed then to avoid waste of time and to provide for the rapid exploration of a subject's possibilities of accomplishment is a series of over-lapping point scales in which a

certain number of tests may be found within the range of every child.

This of course is just another way of describing the Binet scale. The Binet system in which the experimenter finds the basal age and then gives more and more advanced tests until the subject fails several in succession enables the experimenter to give the subject a number of tests within the range of his ability. The four Princeton experimenters using the narrow range method of testing actually averaged about 19 tests for each individual, a number which is remarkably close to the 20 tests prescribed by Yerkes. The proof that this system is reliable rests on the fact that the Binet system has been found to be reliable within certain ranges. It breaks down at the extremes or in those regions below which or above which there are no more effective tests by which the subjects may be differentiated. The chief limit is the number of tests.

So far then our analysis of the nature of an adequate measuring scale has led to the conclusion that this scale must be a point scale on the Binet age scale basis, or in other words that both scales have desirable features. Yerkes' point scale is superior in that it has the partial credit rather than the all or none system of scoring, and the Binet age scale method is superior in that it has greater adaptibility so that the experimenter can find more tests within the subjects' range of ability. The advantage of having post-experimental norms is not peculiar to the point scale for it is just as easy to compute the average "mental age" of groups of non-selected children of different age, sex, nationality, sociological status, etc., as it is to compute the average number of points scored. From a practical standpoint, both scales are probably satisfactory within certain limits. Both scales have a large number of users and have yielded valuable results.

The chief objection against the Binet type of scale is that it is a closed system of tests and admits of no improvement. The moment an experimenter changes a test or the position of a test in the various Binet scales, all the norms of past experimental work are useless. The study of the individual Binet

tests in this investigation certainly shows that the scale can be improved, yet its structure precludes the possibility of improvement without discarding all the norms of previous experimental work. In admitting that the individual tests in the Stanford revision vary in the degree with which they correlate with intelligence (see page 231), Terman admits that the scale can be improved, yet it is impossible to improve it without changing the "intelligence quotients" so carefully worked out on an empirical basis. The chief objection to Yerkes' scale is its lack of adaptability, the fact that its structure places a limit on the number of tests that can be found within the range of ability of the children of the ages for which it is designed to test.

The need of a scale that has an elastic structure and one that may be improved upon indefinitely brings us back to Huey's (36) original conception of a point scale, which consisted of "the per cent. of intelligence obtained by adding together all the points earned, multiplying by 100, and dividing by the sum of the points allotted to the tests actually given and counted as given", or in other words the per cent. that the number of points scored is of the possible number of points. On this basis the norms of performance are not given for any whole system of tests, but for the individual tests. Each experimenter must select his tests to meet his problem, and must compute his norms from the published norms on individual tests.

Such a system has the obvious advantages of adaptability and improvability. If the problem were that of differentiating younger children, the experimenter would select tests that others have found useful in this respect. Under present conditions there is no measuring scale for younger children in which an experimenter can use such an historically valuable test as the form board. If the problem were the diagnosis of the higher grades of defect, the experimenter would select tests useful for making this differentiation. Under such conditions each clinic would probably have its own group of tests, and the only basis of comparability of results would be the individual tests. Huey's system is really nothing more than a percentage system of scoring the results of tests, and the advantage of

the method lies in the fact that a system of tests is more reliable than the individual tests—that a person is more apt to hit a bird with the “spread” barrel than he is with the “choke” barrel.

On this system scales could be constructed on almost any basis, the only limit being the number of standards to which the percentages could be referred. The writer has in mind two types of scales which should prove immediately useful, one an age scale, and the other a scale for feeble-mindedness.

A convenient form of age scale would be a series of tests arranged according to increasing difficulty, the arrangement being made according to any arbitrary criterion such as a score of 50% or 75% at certain ages or combinations of ages. Each test would consist of a number of parts empirically weighted. It would probably be convenient to weigh all tests on some arbitrarily selected scale such as 10 which has the advantages of the decimal system. It would be unnecessary to weigh tests differently for eventually it should be possible to have nothing but valuable tests in the scale. The tests could conveniently be placed in groups according to equality of difficulty, and the score computed from any number of successive groups of tests. The only limit to the number of tests in a group would be the number of tests that could be found of approximately the same difficulty, and any number of groups could be combined.

For purposes of illustration the writer will take five tests for a group and make every successive twenty tests a successive scale. The form of such a scale would be

A ₁	B ₁	C ₁	D ₁	E ₁	F ₁	N ₁
2	2	2	2	2	2	2
3	3	3	3	3	3	3
4	4	4	4	4	4	4
5	5	5	5	5	5	5

Group ABCD would constitute one point scale, group BCDE another point scale, etc. If the norms on the individual tests are accurate the per cent. that an individual scores on scales ABCD, BCDE, CDEF, etc. should refer to the same age, only of course the scale nearest the subject's range of ability would

give the most accurate estimate. A preliminary series of tests such as A_I, B_I, C_I, D_I, etc. could be given to ascertain the subject's probable range. Indeed a preliminary scale that would roughly place the subject near his proper level could probably be arranged for class-room testing. In this way a whole school system could be tested, selecting the doubtful cases on a class-room scale, and giving these cases more and more detailed examinations, the number of tests that can be found being the only limit. Any test that is found to be inaccurate could immediately be thrown out and another substituted, it being a simple matter to recalculate norms.

A series of scales for feeble-mindedness could be arranged without reference to age norms. The scales could be for the three groups, idiots, imbeciles and morons, limiting the groups by some such arbitrary criteria as Binet proposed, viz. that the idiot is one who never acquires spoken language, and the imbecile one who never acquires written language. Illiteracy could be experimentally determined from imbecility by researches on groups of individuals who after long training have not been able to acquire written language. The differentiation of the moron from the normal individual requires further experimental work. Within each group the individuals could scale over a range of 100%, and further study would probably indicate the border-line for the sub-divisions "low", "middle" and "high". These scales could be improved and elaborated as more diagnostic tests were discovered. Of course the scales would not necessarily be limited to the groups for which they were designed. The imbecile could be given the scale for morons, just as well as either one could be given the age scale for normal individuals or as normal individuals could be given the feeble-minded scales. It is the writer's opinion that if scales for feeble-mindedness could be developed that were not standardized on an age basis and the results were then studied in the light of their age correlations, the results would be illuminating.

The chief advantage of systems of tests modelled according to Huey's plan is that they admit of improvement. It is probably impossible for any one person to perfect a final quantitative

scale. It would seem better to adopt a skeleton plan which would allow a more perfect scale to evolve. Such a plan would encourage researches on individual tests both in the public schools and in institutions for the feeble-minded, and it is only by co-ordinating such researches that we will ever be able to solve the problem of the nature of intelligence or the corresponding problem, the nature of feeble-mindedness.

BIBLIOGRAPHY

1. ABELSON, A. R. The Measurement of Mental Ability of "Backward" Children. *Brit. J. of Psychol.*, 1911, 4, 268-314.
2. AYRES, L. P. The Binet-Simon Measuring Scale of Intelligence: Some Criticisms and Suggestions. *Psychol. Clinic*, 1911, 5, 187-196.
3. BATEMAN, W. G. The Naming of Colors by Children. *Ped. Sem.*, 1915, 22, 469-486.
4. BINET, A. Nouvelles recherches sur la mesure du niveau intellectuel chez les enfants d'école. *Année psychol.*, 1911, 17, 145-201.
5. BINET, A. AND SIMON T. Méthodes nouvelles pour le diagnostic du niveau intellectuel des anormaux. *Année psychol.*, 1905, 11, 191-244.
6. BINET, A. AND SIMON T. Application des méthodes nouvelles au diagnostic du niveau intellectuel chez des enfants normaux et anormaux d'hospice et d'école primaire. *Année psychol.*, 1905, 11, 245-336.
7. BINET, A. AND SIMON T. Le développement de l'intelligence chez les enfants. *Année psychol.*, 1908, 14, 1-94.
8. BINET, A. AND SIMON T. La mesure du développement de l'intelligence chez les jeunes enfants. *Bull. de la soc. libre pour l'étude psychol. de l'enfant*. 1911, 11, 187-256.
9. BLOCH, E. AND PREISS, A. Ueber intelligenzprüfungen an normalen Volksschulkindern nach Bobertag. (Methode von Binet und Simon) *Zsch. f. angew. Psychol.*, 1912, 6, 539-547.
10. BOBERTAG, O. Ueber Intelligenzprüfungen (nach der Methode von Binet und Simon). I. Methodik und Ergebnisse der einzelnen Tests. *Zsch. f. angew. Psychol.*, 1911, 5, 105-203. II. Gesamtergebnisse der Methode. *Zsch. f. angew. Psychol.*, 1912, 6, 495-537.
11. BOLTON, T. L. The Growth of Memory in School Children. *Amer. J. of Psychol.*, 1892, 4, 362-380.
12. BONSER, F. G. The Reasoning Ability of Children of the Fourth, Fifth and Sixth School Grades. New York: Columbia Univ., 1910, pp. 133.

13. BRIDGMAN, O. Mental Deficiency and Delinquency. *J. of Amer. Med. Assoc.*, 1913, 61, 471-472.
14. BRIGHAM, C. C. An Experimental Critique of the Binet-Simon Scale. *J. of Educ. Psychol.*, 1914, 5, 439-448.
15. Buffalo conference. J. C. Bell, C. S. Berry, W. S. Cornell, E. A. Doll, J. E. W. Wallin, G. M. Whipple, Informal Conference on the Binet-Simon Scale: Some Suggestions and Recommendations. *J. of Educ. Psychol.*, 1914, 5, 95-100.
16. BURT, C. Experimental Tests of General Intelligence. *Brit. J. of Psychol.*, 1910, 3, 94-177.
17. BURT, C. AND MOORE, R. G. The Mental Differences between the Sexes. *J. of Exp. Ped.*, 1912, 1, 273-284, 355-388.
18. CHOTZEN, F. Die Intelligenzprüfungsmethode von Binet-Simon bei schwachsinnigen Kindern. *Zsch. f. angew. Psychol.*, 1912, 6, 411-494.
19. DECROLY, O. AND DEGAND J. La mesure de l'intelligence chez des enfants normaux d'après les tests de M. Binet et Simon: nouvelle contribution critique. *Arch. de psychol.*, 1910, 9, 81-108.
20. DESCOEUDRES, A. Les tests de Binet et Simon et leur valeur scolaire. *Arch. de psychol.*, 1911, 11, 331-350.
21. DESCOEUDRES, A. Exploration de quelques tests d'intelligence chez des enfants anormaux et arriérés. *Année psychol.*, 1911, 11, 351-375.
22. DOLL, E. A. Inexpert Binet Examiners and their Limitations. *J. of Educ. Psychol.*, 1913, 4, 607-609.
23. DOUGHERTY, M. L. Report on the Binet-Simon Tests given to 483 Children in the Public Schools of Kansas City, Kansas. *J. of Educ. Psychol.*, 1913, 4, 338-352.
24. DRESSLAR, F. B. Studies in the Psychology of Touch. *Amer. J. of Psychol.*, 1894, 6, 313-368.
25. EBBINGHAUS, H. Ueber eine neue Methode zur Prüfung geistigen Fähigkeiten und ihre Anwendung bei Schulkindern. *Zsch. f. Psychol.* 1897, 13, 401-459.
26. FERNALD, W. E. The Diagnosis of the Higher Grades of Mental Defect. *Amer. J. of Insan.*, 1914, 70, 741-752.
27. GILBERT, J. A. Researches on the Mental and Physical Development of School Children. *Stud. fr. Yale Psychol. Lab.*, 1894, 2, 40-100.
28. GODDARD, H. H. The Binet-Simon Measuring Scale for Intelligence. (Revised edition) Vineland, N. J. The Training School, 1911, pp. 16.

29. GODDARD, H. H. Standard Method of giving the Binet Test. Training School, 1913, 10, 23-32.
30. GODDARD, H. H. Two Thousand Normal Children Measured by the Binet Measuring Scale of Intelligence. Ped. Sem., 1911, 18, 232-259.
31. GODDARD, H. H. Three Annual Testings of 400 Feeble-Minded Children and 500 Normal Children. Psychol. Bull. 1913, 10, 75-77.
32. HAINES, T. H. Diagnostic Value of some Performance Tests. Psychol. Rev., 1915, 22, 299-305.
33. HEALY, W. The Individual Delinquent. Boston: Little Brown & Co., pp. 830.
34. HEALY, W. AND FERNALD, G. M. Tests for Practical Mental Classification. Psychol. Monog. 1911, 13 (No. 54) pp. 53.
35. HUEY, E. B. The Binet Scale for Measuring Intelligence and Retardation. J. of Educ. Psychol., 1910, 1, 435-444.
36. HUEY, E. B. A Point Scale of Tests for Intelligence. Baltimore: Warwick & York (folder) 4 pp.
37. KATZENELLENBOGEN, E. W. A Critical Essay on Mental Tests in their Relation to Epilepsy. Epilepsia, 1913, 4, 130-173.
38. KITE, E. S. The Binet-Simon Measuring Scale of Intelligence. Philadelphia: Committee on Provision for the Feeble-Minded, Bull. no. 1, pp. 29.
39. KITE, E. S. The Development of Intelligence in Children. (Contains translations of nos. 5, 6, and 7). Vineland, N. J.: The Training School (Publications of the Department of Research, No. 11), 1916, pp. 328.
40. KITE, E. S. The Intelligence of the Feeble-Minded. (Translation of three articles by Binet and Simon on Feeble-mindedness) Vineland, N. J.: The Training School, (Publications of the Department of Research, No. 12), 1916, pp. 328.
41. KOHS, S. C. The Binet-Simon Measuring Scale of Intelligence: an Annotated Bibliography. J. of Educ. Psychol., 1914, 5, 215-224, 279-290. 335-346.
42. KOHS, S. C. The Practicability of the Binet Scale and the Question of the Borderline Case. Training School, 1916, 12, 211-224.
43. KUHLMAN, F. Some Results of Examining a Thousand Public School Children with a Revision of the Binet-Simon Tests of Intelligence by Untrained Examiners. J. of Psycho-Asthenics, 1914, 18, 233-269.

44. MARTIN, A. L. A Contribution to the Standardization of the De Sanctis Tests. *Training School*, 1916, 13, 93-110.
45. MEUMANN, E. Vorlesungen zur Einführung in die experimentelle Pädagogik und ihre psychologischen Grundlagen. Leipzig: W. Englemann 1913, Vol. II, pp. 800.
46. MEUMANN, E. Ueber eine neue Methode der Intelligenzprüfung und über den Wert der Kombinationsmethoden. *Zsch. f. päd. Psychol. und exp. Päd.*, 1912, 13, 145-163.
47. MORROW, L. AND BRIDGMAN, O. Delinquent Girls Tested by the Binet Scale. *Training School*, 1912, 9, 33-36.
48. NORSWORTHY, N. The Psychology of Mentally Deficient Children. New York: (Columbia Univ. thesis) 1906, pp. 111.
49. OTIS, A. S. Some Logical Aspects of the Binet Scale. *Psychol. Rev.* 1916, 23, 129-152, 165-179.
50. OTIS, M. The Binet Tests Applied to Delinquent Girls. *Psychol. Clinic*, 1913, 7, 127-134.
51. PETERSON, A. M. AND DOLL, E. A. Sensory Discrimination in Normal and Feeble-Minded Children. *Training School*, 1914, 11, 110-118, 135-144.
52. PILLSBURY, W. B. The Psychology of Reasoning. New York: D. Appleton & Co., 1910, pp. 304.
53. PYLE, W. H. A Psychological Study of Bright and Dull Pupils. *J. of Educ. Psychol.*, 1915, 6, 151-156.
54. ROGERS, A. L. AND MCINTYRE, J. L. The Measurement of Intelligence in Children by the Binet-Simon Scale. *Brit. J. of Psychol.*, 1915, 7, 265-299.
55. RUGER, H. A. Sex Differences in the Solution of Mechanical Puzzles. (In report of New York branch of American Psychological Assoc.) *J. of Phil., Psychol., etc.*, 1914, 11, 412-413.
56. SCHMITT, C. The Binet-Simon Tests of Mental Ability. *Ped. Sem.* 1912, 19, 186-200.
57. SCHMITT, C. Standardization of Tests for Defective Children. *Psychol. Monog.*, 1915, 19 (No. 83) pp. 181.
58. SIMPSON, B. R. Correlations of Mental Ability. New York: Columbia Univ., 1912, pp. 122.
59. SMITH, F. O. The Effect of Training in Pitch Discrimination. *Univ. Iowa Stud. in Psychol.*, Vol. VI. *Psychol., Monog.*, 1914, 16 (No. 69) 67-103.
60. STENQUIST, J. L., THORNDIKE, E. L. AND TRABUE, M. R. The Intellectual Status of Children who are Public Charges. *Arch. of Psychol.* 1915. 33, pp. 52.

61. STERN, W. Die differentielle Psychologie in ihren methodischen Grundlagen. Leipzig: Barth, 1911, pp. 503.
62. STERN, W. The Psychological Methods of Testing Intelligence. (Whipple, G. M., trans. fr. German) Educ. Psychol. Monog., No. 13, Baltimore: Warwick & York, 1914, pp. 160.
63. Symposium on Mental Tests. (Conducted by C. E. Seashore under "Communications and Discussions") J. of Educ. Psychol., 1916, 7. (R. M. Yerkes, 163-164).
64. Terman, L. M. Genius and Stupidity. Ped. Sem., 1906, 13, 307-373.
65. Terman, L. M. The Measurement of Intelligence. Boston: Houghton Mifflin Co., 1916, pp. 362.
66. Terman, L. M. AND CHILDS, H. G. A Tentative Revision and Extension of the Binet-Simon Measuring Scale of Intelligence. J. of Educ. Psychol., 1912, 3, 61-74, 133-143, 198-208, 277-289.
67. Terman, L. M., Lyman, G., Ordahl, G., Ordahl, L., Galbreath, N. AND Talbot, W. The Stanford Revision of the Binet-Simon Scale, and some Results from its Application to One Thousand Non-Selected Children. J. of Educ. Psychol., 1915, 6, 551-562.
68. THOMPSON, H. B. Psychological Norms in Men and Women. Chicago: Univ. of Chicago Press, 1903, pp. 188.
69. THORNDIKE, E. L. The Significance of the Binet Mental Ages. Psychol. Clinic, 1914, 8, 185-189.
70. THORNDIKE, E. L. An Introduction to the Theory of Mental and Social Measurements. New York: Teachers' College, 1913, pp. 277.
71. THORNDIKE, E. L., LAY W. AND DEAN, P. R. The Relation of Accuracy in Sensory Discrimination to General Intelligence. Amer. J. of Psychol., 1909, 20, 364-369.
72. TOWN, C. H. A Method of Measuring the Development of The Intelligence of Young Children. (Authorized translation of no. 8) Lincoln, Ill.; Courier-Herald Co. 1913, pp. 82.
73. WALLIN, J. E. W. Experimental Studies of Mental Defectives. Educ. Psychol. Monog. No. 7. Baltimore, Warwick & York, 1912, pp. 155.
74. WITMER, L. On the Relation of Intelligence to Efficiency. Psychol. Clinic, 1915, 9, 61-86.
75. WHIPPLE, G. M. Manual of Mental and Physical Tests. Baltimore: Warwick & York, 1910, pp. 534.

76. WHIPPLE, G. M. *Manual of Mental and Physical Tests*. Baltimore: Warwick & York, 1914, pp. 690, 2 vol.
77. W[HIPPLE], G. M. The Amateur and the Binet-Simon Tests. *J. of Educ. Psychol.*, 1912, 3, 118-119.
78. W[HIPPLE], G. M. Amateruism in Binet Testing once more. *J. of Educ. Psychol.*, 1913, 4, 301-302.
79. WOOLEY, H. T. A New Scale of Mental and Physical Measurements for Adolescents and some of its Uses. *J. of Educ. Psychol.* 1915, 6, 521-550.
80. WOOLEY, H. T. AND FISHER, C. R. Mental and Physical Measurements of Working Children. *Psychol. Monog.* 1914, 18 (No. 77) pp. 247.
81. WYATT, S. The Quantitative Investigation of Higher Mental Processes. *Brit J. of Psychol.*, 1914, 6, 109-133.
82. YERKES, R. M., BRIDGES, J. W. AND HARDWICK, R. S. *A Point Scale of Measuring Mental Ability*. Baltimore: Warwick & York, 1915, pp. 213.

Vol. XXIV
No. 2

PSYCHOLOGICAL REVIEW PUBLICATIONS

Whole No. 103
1917

THE Psychological Monographs

EDITED BY

JAMES ROWLAND ANGELL, UNIVERSITY OF CHICAGO
HOWARD C. WARREN, PRINCETON UNIVERSITY (*Review*)
JOHN B. WATSON, JOHNS HOPKINS UNIVERSITY (*J. of Exp. Psych.*)
SHEPHERD I. FRANZ, GOVT. HOSP. FOR INSANE (*Bulletin*) and
MADISON BENTLEY, UNIVERSITY OF ILLINOIS (*Index*)

Radiometric Apparatus for Use in Psychological and Physiological Optics

Including a Discussion of the Various Types of Instru-
ments that have been used for Measuring
Light Intensities

BY

C. E. FERREE AND GERTRUDE RAND
Bryn Mawr College

PSYCHOLOGICAL REVIEW COMPANY
PRINCETON, N. J.
AND LANCASTER, PA.

AGENTS: G. E. STECHERT & CO., LONDON (2 Star Yard, Carey St., W. C.);
LEIPZIG (Koenigstr., 37); PARIS (16 rue de Condé)



PREFACE

Six years ago, realizing the fundamental relation of energy measurements to quantitative work in psychological optics, we undertook to procure a non-selective radiometer which not only would be sufficiently sensitive for work in the visible spectrum, but the operation of which would be within the technical possibilities of the laboratories in which research is done in psychological optics. At that time we decided upon an instrument of the surface type because (a) we wished to measure all of the light falling on the opening of our campimeter screen rather than compute it from several measurements with the linear type of instrument; and (b) we believed that sensitivity would be added to the instrument in some proportion to the increase in area of the receiving surface (see this paper p. 12). A quick acting surface thermopile, because of its superior steadiness and ease of operation, seemed to be best adapted to our purpose. Since such an instrument could not be obtained in the market, Dr. W. W. Coblentz of the radiometric division of the Bureau of Standards who had at that time just finished a comparative study of the radiometric instruments showing the above-mentioned advantages of the thermopile, undertook the design and construction of the thermopiles, surface and linear, and the auxiliary radiometric apparatus which we have used in our work. These thermopiles are the first of their type made by Dr. Coblentz (see Bulletin of the Bureau of Standards, 1913, 9, pp. 15-29) and have been used by us for five years. Thermopiles of this type are now in use also in many physical laboratories and in other laboratories in which a sensitive and convenient means is needed for measuring spectrum energies.* We

*For example, the recent work of Nutting (*Phil. Mag.*, 1915, 29, (6), p. 301), Ives, Coblentz and Kingsbury (*Phys. Rev.*, 1915, 5, (2), p. 269), and Coblentz and Emerson (*Bull. Bur. of Standards*, 1917, 14, p. 167), on the visibility of radiation and the mechanical equivalent of light, of interest to psychologists, has been done and was made conveniently possible by these improved thermopiles designed and constructed by Dr. Coblentz.

are not thus recommending in the following pages an apparatus the feasibility and convenience of which for quantitative work of the kind needed in psychological optics is untried.

Of late some dispute seems to have arisen with regard to the need and uses of energy measurements for work in psychological optics. A brief discussion and statement of opinion on this point was given by us in an article published in the *American Journal of Psychology* in 1912. (A Note on the Determination of the Retina's Sensitivity to Colored Light in Radiometric Units, 23, pp. 328-332.) Time alone can of course reveal the full range of needs and uses of this type of measurement and render a just verdict on disputed points. A few words in the way of general perspective, however, may not be out of place here.

Considered in its relation to the eye two points of view may be recognized in the rating of lights. One of these is involved in their rating for the use of the eye as an organ of seeing. In such a rating it is obvious that the method used should take into account all of the eye's deviations from equality of response to the different wave-lengths of light. In the production of illuminating effects this is the work of photometry which should be done by the eye or some instrument calibrated to give results in terms of the responses of the eye. Another and quite different point of view, however, is involved in their rating for the purpose of investigating the eye's peculiarities or characteristics of response in every way in which it is capable of giving response. In such work it is obvious that the ultimate method of making the rating should be free from the peculiarities to be investigated, that is, should not be made by the eye itself. In general in work of this kind two needs arise. (1) A method of specification is required that will make possible an accurate and convenient reproduction of intensities from time to time and from laboratory to laboratory. The difficulty of doing this by photometry with lights differing widely as to wave-length as do in most cases the stimuli employed in psychological optics, is too well known to need emphasizing here. Obviously what is needed for certitude in this work is a measuring instrument

which can be calibrated directly against the standard of radiation, or black body, and which is non-selective in its response to wave-length,—not an instrument like the eye, the selenium cell, the photo-electric cell, or the photographic plate, the responses of which are not only selective to wave-length but vary in their amounts of selectiveness with change of intensity of light, differ greatly in both of these regards (especially the eye) from instrument to instrument or from sense organ to sense organ, and can be calibrated against the radiation standard or black body, if at all, only with a great deal of difficulty and with many chances of cumulative error.

The insistence on a subjective method of rating for standardizing purposes when the objective method is available, is not only difficult to understand but is entirely contradictory to current practice in other sense fields. No one would think, for example, of specifying for the purpose of securing reproducibility, the weights used in an investigation of skin sensitivities in terms of the skin's own responses when the means of making the physical measurement is at hand; yet there should be more chance of successfully establishing from laboratory to laboratory a system of calibration of skin measurements in terms of some common standard than there is of accomplishing the analogous task in case of light. It is scarcely conceivable that the most ardent advocate of subjective ratings in case of light would recommend the substitution of the subjective for the objective method for work on the skin for the simple reason that it would be so undesirable as not to be tolerated, unless for want of an objective method it was rendered absolutely necessary. With the objective method available from the beginning, the possibility of using the subjective method has not even been raised in work on the skin. And indeed the subjective method has been used in rating light intensities only because (a) for more than a hundred years no other method was available, and (b) it was desirable to rate lights for use in seeing by a method which gave results corresponding to the eye's powers of response. The former of these reasons for its use has now disappeared. Only the latter, with a few laboratory exceptions, remains and marks

off for the subjective method of rating, a separate and special field which is clearly recognized as such by physicists and the engineers dealing with the problem of lighting.

As a brief, however, for the continuation of the use of the eye for the measurement of its own stimuli, although such measurements would not be subjective, it may be claimed that in time it will be possible to calibrate the eye by means of the non-selective radiometers so that it can be used to measure the visible energies directly. For example, just as it is possible to measure a linear dimension with a meter rod and to convert the results into terms of the English system, and *vice versa*; so it may be possible to measure the different wave-lengths of light by the eye and convert the results obtained into energy values. The difficulties in the way of this, as we have already pointed out, consist in differences in the sensitivity of different eyes for a given wave-length; the selectiveness of the eye's response to wave-length and to intensity and its variations in both of these regards from observer to observer; the lack of a fixed scale from observation to observation, even in the case of a single observer; etc. In short to complete the analogy suggested above, it would be an exceedingly difficult task to convert measurements from the metric into the English system and *vice versa* if very few of the measuring rods employed represented the same amounts of linear space; if in case of a given rod the dimensions of some objects were over-estimated and others under-estimated and the magnitude of this over-estimation and under-estimation varied with the dimensions of the object measured by amounts as yet undetermined; etc.,—as happens in case of the eye's evaluations of the wave-lengths of the visible spectrum. Obviously if the eye's ratings are to be converted into energy values, it can come only after a very great deal of investigation and calibration against radiation standards by means, for example, of the non-selective radiometers, which but constitutes one of the subdivisions of what we have included under the second of the needs we are giving for energy measurements in the study of the responses of the eye. However, to represent the calibration as now completed and available for use instead of scarcely begun,

would be chimerical and visionary to a degree which we can consider compatible only with an insufficient knowledge and understanding of all that is involved in the problem.

Since the foregoing was written, Troland (*Journal of Experimental Psychology*, 1917, 2, pp. 7-13) has advised that, instead of the thermopile or other non-selective radiometer, the psychologist may, with sufficient accuracy for his purpose, use the eye as a selective radiometer and convert the results into units of energy by means of a value for the mechanical equivalent of light that has recently been determined by Nutting (*Philos. Mag.*, 1915, 29, (6), p. 301). This point can not be discussed here in detail. We would, however, recommend that the reader consult this work on the mechanical equivalent, which has been done by means of the flicker photometer and the thermopile, and judge for himself how unreliable it would be to attempt to follow Dr. Troland's advice and use a result obtained with a given limited group of observers for only one intensity of light, to convert the photometric results of individual observers in other laboratories and for other intensities of light into anything at all closely approximating the correct energy values. It is obvious that in order to make the conversion in any given case with the same order of accuracy with which the direct energy measurements may be made, the same observers would have to be used, the same state of adaptation and sensitivity of the eye, the same intensity of light or approximately so (at least so far as adequate proof to the contrary for a large part of the intensity scale is concerned), the exact same range of wave-lengths and distribution of energy within the group of wave-lengths, and the same degree of purity of light as were used in making the original determination of the visibility curve which is meant to serve as the basis for making the conversion. Considering the first of these points alone, it will be remembered that Ives, working through the spectrum with the flicker photometer, found in a group of eighteen observers disagreements as great as 159 per cent for $.487\mu$; 114 per cent for $.498\mu$; 26 per cent for $.518\mu$; 18 per cent for $.537\mu$; 13 per cent for $.556\mu$; 10 per cent for $.576\mu$; 28 per cent for $.595\mu$; 65 per cent for $.615\mu$; 86 per cent for $.635\mu$; and 122 per cent for $.655\mu$. (*Philos. Mag.*, 1912, 24, Ser. 6, pp. 856-863.) From this showing of low agreement from observer to observer with the flicker photometer, it is clear that the results for individual observers could not be used for the purpose of making the conversions recommended unless some means were had of correcting these results to those of the group for which the original visibility curve and the mechanical equivalent were determined. Space can not be taken here to discuss the complications and approximations that would be involved in making such a correction. It will be sufficient to say that if it were made in the most approved manner—an adequate method of doing it has not by any means as yet been devised—and the mechanical equivalent were applied, the results would scarcely be accepted as correct even by an ordinarily careful worker unless they could be checked up by a direct energy measurement. In this connection it is interesting and important to note that the visibility curve as determined by Nutting does not agree with that determined by Ives, also that the curves of Nutting and Coblentz agree only when certain

corrections are made in the energy measurements of Nutting. In short the attempt to get a set of figures that will express for the different wave-lengths the relation of the lumen as evaluated by a number of eyes to the watt is an interesting bit of work and may present, perhaps, when the proper computations are made, a rough analogy to the determination of the mechanical equivalent of heat; but the attempt to use these figures to convert the photometric results of the individual observers in the different laboratories into the correct energy values is quite a different matter, and can scarcely be considered as the intent of those who have made the determination. This question will be discussed in greater detail in a later paper.

However, the idea of using the eye indirectly to determine the energy values of light is by no means new. Before the direct type of measurement had been made as feasible as it now is, several attempts were made to use the eye for this purpose. (See, for example, Lummer and Pringsheim. *Jahresber. d. Schles. Ges. f. vaterl. Kultur*, 1906, pp. 95-97; Beibl., 1907, p. 466; Thürmel, *Das Lummer-Pringsheimsche Spektral-Flickerphotometer als optisches Pyrometer*, *Ann. der Phys.*, 1910, 33, (4), pp. 1139, 1160; etc.)

(2) The second and perhaps more fundamental need for energy measurements for work on the eye is, as stated in the general heading, for a method of rating the stimulus which will make possible a quantitative comparison of the eye's power of response to its stimuli in every way in which it is capable of giving a response; for we can know the kind and amount of its selectiveness of reaction to the different wave-lengths of light only when they are compared with those of an instrument as standard which shows equal power or capacity of response to all wave-lengths. Only with such an instrument, or rather with such an evaluation of the stimuli as a common or invariable standard to which to refer the eye's evaluations or responses, can the work of comparing its powers or peculiarities of response to its stimuli be put on a basis that can be called quantitative for a single eye or from eye to eye. To this it may be demurred, however, that in some problems it is required as one of the features of the investigation that the stimuli have equal power to arouse the eye's response or sustain some subjective relation to each other. This need we have always freely recognized both in our work and in our recommendations. (See *Amer. Jour. of Psychol.*, 1912, 23, pp. 328-332.) It in no way conflicts with, however, or supplants the more fundamental one already given, but is rather supplementary to it in certain types

of investigation; for even in the cases where the subjective relation is demanded to fulfill the requirements of the investigation there is still great need for the ultimate purposes of the science that the physical amounts of light required to produce this subjective relation for the given observer be determined and specified. Again to use the analogy of work on skin sensation, it would be a careless investigator indeed who would fail to specify, if it were possible to do so, the physical measure of the weights he used to give, for example, equal pressure responses. In short, it seems a paradox that one should even feel the need to make a special pleading for the introduction of objective measurements into the work of psychological optics when it is the current practice to use objective ratings of the stimulus in every other field of psychological investigation in which it is possible to do so, the intensity ratings in vision and audition alone being the conspicuous outstanding exceptions and these being so only because adequate methods for making such ratings have been slow in coming.

As examples of needs for regulating the stimuli to give certain subjective relations we may quote here the following cases that we have already formally recognized. In a recent investigation of the comparative lags of the achromatic response to wavelength made in our laboratory, the stimuli employed were made photometrically equal and the amounts of light used to give these equal responses were measured radiometrically. The photometric equalizations were made because the data were wanted in an interpretation of the characteristic overestimations and underestimations found in the results of certain observers in photometry by the method of flicker as compared with their results by the equality of brightness method. In another investigation now in progress, namely, a determination of whether stimuli which have the same power to arouse the achromatic response have also the same power to make the eye lose in its capacity to give this response as a result of prolonged stimulation, the stimuli are as a matter of course being made subjectively equal as one of the essential conditions of the investigation; but again the amounts of light required to produce this subjective re-

lation will be determined radiometrically for the purpose of ultimate specification. Also in our original note on energy measurements we recognized quite broadly the possible need of establishing subjective relations between the stimuli used. For example, on p. 329 in discussing methods of determining after-image and contrast sensitivity, we state: "It is conceivable that two points of view may be held with regard to what is meant by after-image and contrast sensitivity. (1) After-image and contrast sensitivity may express a relation between the amounts of light required to arouse after-image and contrast sensations and the unit of light used. (2) It may express a relation between the amount of light required to arouse the after-image and contrast sensations and the amount required to arouse the positive sensation." In the former case the after-image or contrast sensations are treated as one of the eye's responses the selectiveness of which to wave-length is to be determined; in the latter a figure is sought which expresses the relation between the after-image and contrast and the positive sensitivities. On the same and the succeeding page we say: "Similarly two views may be held with regard to the determination of the comparative rates of fatigue, and of the development-time of sensation. (1) Lights equalized in energy may be used. (2) The energy of the lights may be made proportional to the sensitivity of the eye to the different colors." Also in discussing the investigation of the peripheral limits of sensitivity, we state: "(a) The limits may be considered in relation to the comparative sensitivity of the retina to the different colors. (b) They may be considered in relation to existing color theories. In the first of these problems the limits should be obtained with stimuli equalized in energy. So obtained the results will constitute merely another expression of the comparative sensitivity of the retina to the different colors. The second problem is more complicated and will be made the subject of a separate paper." Indeed as these citations abundantly show, we have never failed to recognize that the stimuli in certain types of investigation must be made to conform to some type of subjective relation, but these investigations constitute in immediate importance only a minor part of the

work that is to be done in getting a thorough knowledge of the eye's characteristics of response; and even in these investigations there is as great need of an invariable standard of reference as there is in any field, psychological or otherwise, where the value of quantitative work or measurement is recognized.

Perhaps the general character of the discussion will not be deviated from too widely if we add in conclusion a word on the determination of retinal sensitivities which will indicate in a concrete case the type of treatment that should in our opinion be given both to the response and to the stimulus, when possible, in quantitative work in psychological optics. If the sensitivity of the retina is to be measured in a way that is comparable with the measurement of the sensitivity of the physical recording instruments, two conditions must be fulfilled: (a) the amounts of response in terms of which the comparison is to be made must be numerically comparable; and (b) the amounts of stimulus used in arousing the response must also be numerically comparable, or commensurable. The sensitivity of two galvanometers could not be compared, for example, were it not known that the divisions on the scale of each were either equal or commensurable; likewise the amounts of current used to produce the given deflections must be known in terms of the same, or comparable units. With the introduction of the radiometric treatment of the stimulus the second of the above conditions is fulfilled, and for the first time in a way that can be considered as quantitative to a degree that would be acceptable in the rating of the sensitivity of a physical instrument. With reference to the first condition we are confronted with a situation somewhat similar to that which obtains in heterochromatic photometry. That is, in general five different quantities have been used or suggested in the work of measuring sensitivities (the liminal threshold, the just noticeable difference, the average error, equal amounts of response and equal sense differences), but only the last two of these, so far as has yet been demonstrated, conform with certainty to the requirement that is considered absolutely necessary in determining the sensitivity of a physical instrument, namely, that the amounts of response as well as the amounts of

stimulus must be numerically comparable. Moreover, in the absence of sureness of principle in case of the other three, the empirical check of agreement in result with those that have the needed sureness of principle has never been offered; yet sensitivities are determined just as if this condition did not exist, comparisons are made and conclusions are drawn. In short it may not be out of place to call attention here to the looseness of thinking and practice that prevails more or less generally with regard to the work of determining physiological sensitivities as compared with the analogous physical determinations. For the sake of consistency it might well be urged either that this work be revised on the basis of the standards set for the physical instruments with all of the interchecking of methods that is needed, or that the term sensitivity with its definite quantitative connotation be abandoned in all cases in which this standard can not be lived up to.

TABLE OF CONTENTS

	PAGE
I. Introduction	I
II. Methods and apparatus that have been used for the measurement of light intensities and their applicability to the investigation of retinal sensitivities....	2
A. The thermopile	9
1. Important points in the construction of sensitive thermopiles	10
a. The metals used to form the thermoelectric junctions	10
b. The size of the wire used in forming the couples	11
c. The dimensions of the pile and the number and arrangement of the thermocouples	12
d. The type of connection of the couples	13
e. The relation of internal to external resistance	13
f. Nicety of construction	14
g. Provisions to secure steadiness of response	15
2. Advantages of the thermopile	15
B. The Nichols radiometer.....	16
1. Significant points with regard to the radiometer	17
2. Comparative advantages and disadvantages of the radiometer	17
C. The radio-micrometer	18
D. The bolometer	20
1. Important points in the construction of sensitive bolometers	21
a. The kind of material used for the receiving surface	21
b. The area of the receiving surface....	21
c. The thickness of the strip used as a receiver	21
d. The most favorable resistance of bolometer and balancing coils	22

	PAGE
e. The slide-wire for balancing the resistance	22
f. The protection of the bolometer from air currents	22
g. The strength of current	23
2. Points in the construction of the auxiliary galvanometer	23
3. Possible sources of difficulty in the use of the bolometer	25
4. The comparative advantages and disadvantages of the bolometer	25
E. The selenium cell	26
1. Points to be considered in the construction and use of selenium cells	27
a. Method of preparing the selenium....	27
b. Purity of the selenium	28
c. Material and size of the electrodes...	28
d. Strength of the battery current.....	29
e. Direction of the battery current.....	30
f. Duration of the battery current.....	30
g. Temperature at which cell is operated	30
h. Pressure	31
i. Moisture	31
j. Age of cell	31
k. The amount of polarization gradually set up in the cell.....	32
1. Photo-electric currents	32
2. Factors which render it difficult to use the selenium cell for quantitative work either as an ohmic resistance or as a light-measuring instrument	32
3. Factors which apply especially to its use as a light-measuring instrument	33
a. The preexposure of the cell.....	33
b. The time of exposure to the light to be measured	34
c. Wave-length of spectrum light	34
(1) Factors which have been found to influence the selectiveness of response to wave-length for a given cell	35
(a) Intensity	35

	PAGE
(b) Length of time of exposure	35
(c) Temperature, humidity and voltage	36
(d) Photo-electric currents..	36
(2) Factors which have been found to influence the selectiveness of response to wave-length in different cells	36
d. Intensity of mixed or white light....	37
4. Theories of the action of light on selenium..	39
F. The photo-electric cell	41
1. Factors that have been taken into account in the construction and use of photo-electric cells	42
a. The metal used to hold the negative charge	42
b. The residual gas	42
c. The pressure of the residual gas.....	43
d. The potential difference between anode and cathode	43
e. The galvanometer or electrometer employed	43
f. The angle of the incident light.....	44
g. Dark effects and after-effects.....	44
h. Fatigue effects	44
2. Comparative advantages and disadvantages of the photo-electric cell	45
G. The photographic plate	47
1. The blackening of the plate and the factors which influence this action	47
2. The possibilities of using the photographic plate in quantitative work	52
H. The eye	52
1. The two possibilities of using the eye in light measurements	52
2. The comparative advantages and disadvantages of the use of the eye for balancing energies of light of the same spectro-radiometric composition	53
III. A convenient and sensitive radiometric apparatus for work in psychological and physiological optics	54

	PAGE
A. The linear thermopile	55
B. The surface thermopile	57
C. The radiation standard	59
D. The galvanometer	61
E. The auxiliary apparatus	63
1. Sensitivity tester for the galvanometer	63
2. Special resistance coils	64
3. The telescope and scale	64

I. INTRODUCTION

In a previous paper the purpose has been expressed of describing apparatus for work on the color sensitivity of the retina consisting of spectroscopic and radiometric features. In partial fulfillment of this purpose apparatus was described in a recent number of the *Journal of Experimental Psychology*¹ designed to meet the following needs: (1) to stimulate any part of the retina with the light of the spectrum and to control as desired the conditions of preexposure and surrounding field; and (2) to regulate the amounts of light used within the small gradations needed for threshold and just noticeable difference determinations. It is the purpose of the present paper to describe apparatus with which it is possible to specify the amount of light used in energy units. This completes the description of a group of apparatus by means of which it is possible to determine the sensitivity of the eye to wave-length in terms that are commensurable, and thus to place the investigation of the responses of the eye on a methodological plane comparable with the study of the responses of the physical recording instruments. This could not be done until a means was had of estimating light intensities which is not only independent of the achromatic and chromatic functioning of the eye itself, but which gives results directly proportional to the physical value or energy of the light waves. An instrument which gives responses directly proportional to the intensity of the light-waves is, we scarcely need to point out, equally sensitive to all wave-lengths. With the responses of such an apparatus as standard, the deviations of the eye from equal sensitivity to the different wave-lengths can readily be determined and compared.

¹ Ferree and Rand. A Spectroscopic Apparatus for the Investigation of the Color Sensitivity of the Retina, Central and Peripheral. *J. of Experimental Psychology*, 1916, 1, pp. 247-283.

II. METHODS AND APPARATUS THAT HAVE BEEN USED FOR THE MEASUREMENT OF LIGHT IN- TENSITIES AND THEIR APPLICABILITY TO THE INVESTIGATION OF RETINAL SENSITIVITIES

No problem in optics probably has presented more difficulty to the investigator and the various committees which have been appointed for the purpose by scientific and engineering societies, bureaus, etc., than that of standardizing the intensity of lights differing in color value. In the investigation of retinal sensitivities the problem of standardizing presents two aspects. (1) As the prime requisite of scientific work a method of specification is needed that will make possible an accurate and convenient reproduction of light intensities. Without this no investigation can have the certitude that comes from repetition and verification. And (2) an important item in the determination of retinal sensitivities has been a comparison of the sensitivity to different wavelengths. This has been made a feature of the general problem for the sake both of knowing the characteristics of the eye as a sense-organ and measuring instrument, and of being able to meet the many practical needs that have arisen in the attempt most effectively to adapt light to the service of the eye in the work of lighting, etc. As we have already pointed out, if the sensitivity or responsiveness of the eye to lights of different wave-lengths is to be compared, it is obvious that a common unit must be had, independent of the functioning of the eye itself, in terms of which to measure the quantity or intensity of the light employed; or to express the need in another form, the light to be used in the investigation should be rated by instruments whose responses are directly proportional to the energy value of the light waves.¹ Such instruments being non-selective in their response to wave-

¹ For a further statement of the conditions that must be fulfilled if the sensitivity of the retina is to be measured in a way that is comparable with the measurement of the sensitivity of the physical recording instruments, see Preface, p. xi.

length and giving the true physical value of the stimulus are the logical standard of reference in the comparative study of instruments or organisms whose responses are selective. Fortunately instruments for measuring light intensities which fulfill the above requirements have within the last few years reached such a stage of advancement as to make this kind of treatment of the problem not only possible but feasible and even convenient. On this account a brief history may not be out of place here of the attempts that have been made to attain measurements of light intensities which are purely physical.

Light of any wave-length is universally conceded to be a form of motion in a transmitting medium. By common agreement among physicists quantitative estimates of motion are made in terms of what is called energy of motion; or of mass and rate of motion. Owing to the small quantities of energy involved in the waves of the visible spectrum, it is obvious that light energies can not be estimated directly in terms of mass and rate of motion. Some instrument or apparatus which responds to light must be used and the response of this instrument be calibrated against a source of energy the radiation from which per unit of surface per unit of time is known. Once calibrated, such an instrument with proper checks on its sensitivity may be used for the measurement of the visible radiations from any source. The requisites of a satisfactory instrument for the physical measurement of light are obviously as follows. (1) It must give a response which is directly proportional to the energy of the light wave or must be capable of calibration against an instrument which does give such responses. (2) It should be non-selective in its response to wave-length and to intensity, *i.e.*, it should be no more sensitive to one wave-length of the spectrum than to another and its sensitivity should not vary with the intensity of the light used. This requirement is an obvious corollary to the preceding. If an instrument is used which is selective in its response to wave-length, the amount of its selectiveness must be a constant else correction factors cannot be determined which will be valid for all intensities. (3) It should be sufficiently sensitive to respond to the small amounts of light present in the visible spectrum. And (4) it

should give results which have a satisfactory degree of reproducibility, or if erratic within known limits or conditions it must be calibrated against some instrument which does give reproducible results, and correction factors be determined.

As comparators of light intensities the following instruments and the human eye have at various times been employed or investigated,—the Nichol's radiometer, the radio-micrometer, the micro-radiometer, the bolometer, the thermopile, the selenium cell, the various types of photo-electric cell, and the photographic plate.² The following comparisons may be made of these instruments with regard to the above mentioned requirements. (a) The radiometer, the radio-micrometer, the micro-radiometer, the bolometer, and the thermopile, depending initially for their action on heating effects, give responses which are directly proportional to the energy of the incident light. They are, therefore, non-selective in their reaction both to wave-length and to intensity.³ The selenium cell, the photo-electric cell, the photographic plate and the human eye, however, do not give responses which are proportional to the energy of the incident light. They are all known to be selective in their reaction to wave-length; and the amount of this selectiveness in case of the selenium cell, the photographic plate and the human eye has been found to change with the intensity of light. (b) All the instruments which are selective in their responses to wave-length, namely, the human eye, the selenium cell, the photo-electric cell and the photographic plate, have a high degree of sensitivity to light. The photographic plate possesses the additional advantage that the action may be integrated over an interval of time. The instruments which are non-selective to wave-length are as a class less sensitive to light. Recent improvements in the construction of such instruments,

² The use of these instruments for the measurement of light is based on the following effects produced by incident light: (1) heating effects; (2) a change in the resistance of certain metals to the flow of a current; (3) a decrease in the power of certain metals to hold a negative charge in a partial vacuum; (4) chemical action; and (5) visual sensation,—used chiefly in connection with the various types of photometer.

³ It should be mentioned, however, that their windows absorb selectively some of the wave-lengths of the invisible spectrum.

however, have increased their sensitivity greatly. Of this class of instruments, the thermopile because of its greater ease of control and greater reliability is probably best adapted for use in laboratories of physiological and psychological optics. Moreover, as was stated in our introduction, it has been developed to a high degree of sensitivity. In fact a comparative study of the non-selective instruments has shown that in the present stage of development of such instruments, the thermopile possesses as high a sensitivity as the others when operated in air and probably also when operated in a vacuum. (c) The factors which influence the response and use of the instruments which are selective in their action to wave-length have proven to be so hard to control that the results obtained have shown a comparatively low degree of reproducibility. Of the non-selective instruments the bolometer is perhaps the hardest to control. The factors which influence the action of the thermopile, the radiometer, and the radio-micrometer are on the other hand comparatively easy to control. A comparative statement of the advantages and disadvantages of these instruments will be given later in the paper.

Before considering these instruments in greater detail, it may be of service perhaps to give a brief statement of the type of action that is produced in each by the radiant energy falling upon its receiving surface. As was stated earlier in the discussion, the measurement of energy by the types of instrument that are mentioned here is not direct. The instrument is available because it gives to a greater or lesser degree some regular and constant type of response to radiant energy, the value of which in energy units is determined by calibration against the known radiations from a black body, or from some other source whose radiations have been determined by comparison with that from a black body. In instruments of the type of the radiometer, micro-radiometer, radio-micrometer, bolometer and thermopile, all radiations are transformed into heat at the receiving surface of the instrument. In case of the radiometer, for example, the absorbed energy produces thermodynamic effects in the rarified gas contained in the housing of the instrument, which in turn causes regular deflections of a delicately suspended vane; in case of the radio-microm-

eter and the thermopile, the absorbed energy acting upon a thermo-electric couple causes a flow of current which deflects the needle of a sensitive galvanometer in circuit with it; and in case of the micro-radiometer and bolometer, the absorbed energy changes the resistance to the flow of current in a delicately balanced electric circuit which is also detected by means of a sensitive galvanometer. The action of the remaining instruments is not due to heating effects; also these instruments are not responsive to all radiations. The selenium cell and the eye, for example, are sensitive only to the visible spectrum (Brown and Sieg, however, *Phys. Rev.*, 1914, 4, (2), pp. 48-61, report one cell that has considerable sensitivity as far out as $.85\mu$.) the photographic plate, when properly sensitized to red, and the photo-electric cell are sensitive both to the visible and the ultra-violet radiations.

In case of the selenium cell the visible radiations falling on a strip of metallic selenium placed in one arm of a delicately balanced electric circuit so change the resistance of the selenium to the flow of current that the electromotive balance between the two arms of the circuit is disturbed, and a flow of current takes place between two given points which were before at equal potentials. This current deflects a galvanometer. The use of the selenium cell is attended with a great deal of difficulty and there are many opportunities for cumulative error. The following is a brief statement of some of these difficulties. A detailed statement will be given later in the paper. (1) As an instrument to be used in the process of measuring, it can be employed without calibration (*e.g.*, the determination of sensitivity curves for different intensities of light expressing a relation between response and energy) only to identify equal amounts of energy; and since it is as a general case responsive only to light waves it can be used to equalize only light energies. Its employment in this way as a measuring instrument for the visible spectrum energies presupposes, therefore, a standard light source, the energy values of the visible spectrum from which are known, against which to balance the unknown lights. But as stated earlier in the paper, the light energy emitted from sources ordinarily available can not be determined directly. It must be determined by compari-

son with the radiations from some body the amount of which can be directly estimated. This comparison may be most conveniently made by means of some measuring instrument such as the thermopile which is responsive to the total of radiation and which is non-selective in its response to wave-length, and a black body radiating known amounts of energy to furnish the standard for the comparison. In short, without the possibility of ultimate recourse to such instruments as the thermopile, the radio-micro-meter, etc., which are non-selective in their response to wave-length, instruments of the class of the selenium cell would be practically useless for radiometric purposes. Moreover, the two-fold nature of the measuring operation, the difficulty of maintaining constancy of conditions in the employment of the secondary standard, and more especially the many factors extraneous to light which influence its response, make its use very liable to error. And (2) since the selenium cell is selective in its responsiveness to the different wave-lengths of light, the standard light source must in every case be of the same spectro-radiometric composition as the light against which it is to be balanced, otherwise the cell can not be relied upon to give to the unknown light a fair radiometric evaluation. That is, if the light to be measured is not of the same wave-length or composition as the standard light, correction factors have to be used which represent the amount of the selectiveness of action. Furthermore, the amount of selectiveness of the action changes with the intensity of light, therefore correction factors established for one intensity will not serve for all intensities.

The action of the photo-electric cell depends on the power of light to cause certain metals to lose a negative charge of electricity in a partial vacuum. Much that has just been said of the selenium cell applies also to the photo-electric cell. (1) It is not sensitive to the infra-red spectrum, hence can not be calibrated directly against the total of radiation of a black body. (2) It is selective in its response to the different wave-lengths of the visible spectrum. Griffith and Dember claim that it is also selective to intensity. (3) Its use in measuring the energy of the visible spectrum presupposes, for example, either some calibration similar

to that noted above for the selenium cell, or the availability of a light source the values of the visible radiations of which are known to serve as a standard against which to balance the unknown wave-lengths. And (4) its sensitivity is influenced by so many factors difficult of control as to give it a comparatively low reproducibility of response.

The photographic plate responds to light by a chemical change in its sensitive film, known as the 'blackening' of the plate. Its convenient use as an energy measuring instrument depends upon whether or not this blackening sustains any constant relation to the amount of incident light. If not, its use would necessitate such an elaborate calibration as to render it impracticable as a radiometer. Like the selenium and photo-electric cells, it too is selective both to wave-length and to intensity; its employment as an energy measuring instrument presupposes a standard light source, the energy value of the radiations from which is known; and its responses are subject to the influence of many variable factors which tend to give them a low degree of reproducibility.

The eye gives two responses to light waves, the chromatic and the achromatic. As yet the achromatic response alone has been used in the measurement of light intensities. Two possibilities are presented for the use of the eye as a measuring instrument: photometric or the rating of lights in terms of their power to arouse the achromatic sensation; and radiometric in the sense of balancing or equalizing the energy values of lights of the same spectro-radiometric composition. As an energy comparator the eye is like the selenium cell in the following regards. (1) It is responsive only to the visible spectrum. Its employment, therefore, presupposes the provision of a light source, the energy value of the visible radiations from which are known. And (2) since it is selective in its response to wave-length, it can without correction factors be used to establish an energy balance only between lights having the same spectro-radiometric composition. While not generally used or classed as radiometric, the eye can like the selenium and photo-electric cells be used very sensitively to balance energies of lights of the same spectro-radiometric

composition and has in this respect a similar claim to be considered as one of the radiometric possibilities. In fact our control of the factors which influence the response of the eye is perhaps enough greater than that of the selenium cell, the photo-electric cell, etc., to render its use for this purpose preferable from the standpoint of precision.

A. THE THERMOPILE.⁴

The thermopile is probably the most celebrated of the radiometric instruments. To it we are indebted for the researches of Melloni and Tyndall as well as for the most notable advances that have been made in the study of radiation. The instrument was invented by Nobili and is based on a discovery made by Seebeck about 1820 that when two wires of different metals are joined end to end so as to form a closed circuit, an electric current passes around the circuit when one of the junctions is heated or cooled, and this current continues to flow as long as any difference of temperature exists between the two junctions.⁵

⁴ With regard to the non-selective radiometers we are indebted heavily to the publications of Dr. Coblentz for data and for guidance in the compilation of data.

⁵ There are three thermo-electric effects in metals: the Seebeck effect, the Peltier effect, and the Thomson effect. The Seebeck effect is described above and is the one on which the action of the thermopile is based. The Peltier effect discovered in 1834 is the converse of the Seebeck effect, *i.e.*, when a current is passed through a junction of dissimilar metals, the junction is either heated or cooled depending upon the direction of the current with reference to the thermo-electric relation of the metals. For example, if the current passes from the electro-negative to the electro-positive, work is done and the temperature of the junction is raised; but if it passes from the electro-positive to the electro-negative, the temperature of the junction is lowered. The result of the Peltier effect in a thermo-couple, therefore, is to lower the temperature of the exposed junction. This effect, however, is not considered to be sufficient to make an appreciable change in the results gotten with a thermopile-galvanometer combination of the sensitivity ordinarily obtained. The Thomson effect is a heat effect manifested when a current flows between points at different temperatures in the same metal. This effect differs in different metals. For example, when a current flows from a hot to a cold point in copper, it evolves heat; but when it flows from a cold to a hot point, heat is absorbed. In iron, however, the reverse is true. When the current flows from a hot to a cold point, heat is absorbed. This effect for the small temperature differences involved is also considered negligible by Altenkirch (*Phys. Zeit.*, 1909, 10, p. 560) in his discussion of the efficiency of thermopiles.

Like the bolometer the thermopile owes its effective sensitivity in part to its own construction and in part to the auxiliary galvanometer.

1. *Important points in the construction of sensitive thermopiles.* The problem in thermopile construction appears to be to secure a low resistance, a low heat capacity and heat conductivity, and a high thermo-electric power. The following have been found to be important points in the construction of thermopiles. a. *The metals used to form the thermo-electric junctions.* This point is of importance because metals are found to differ in their thermo-electric power, *i.e.*, in their electromotive force per degree centigrade when compared with the standard metal, lead. The following are some of the thermo-electric metals: bismuth, silver, German silver, lead, platinum, copper, zinc, iron, antimony, constantan, tellurium, and selenium. A very small amount of impurity may make a great difference in the thermo-electric power of a metal, and some of the alloys and metallic sulphides show a very high thermo-electric power. Some of the combinations most commonly used in making thermo-couples are bismuth and antimony, iron and constantan, and bismuth and silver. The bismuth and silver couple has been chosen by Coblentz because of its high thermo-electric power and low resistance. Silver was selected to complete the element with bismuth more especially because of its low resistance, its pliability and the ease with which it can be cleaned⁶ and annealed. The latter two points are of great importance in the construction of the pile. Nicety of construction is of fact of greater importance to a high radiation sensitivity, Coblentz declares, than a high thermal E. M. F.⁷ provided the material has a correspondingly high resistance.

⁶ It is important that the metal chosen be easily cleaned for completeness of contact in soldering. A preliminary heating can be given the silver wire which serves the double purpose of cleaning and annealing. This preliminary heating could not, for example, be given to copper and iron wire.

⁷ Coblentz (Bulletin of Bureau of Standards, 1914, 11, pp. 148-150) found, for example, in eight samples of bismuth wire with diameters of 0.06, 0.08, 0.1 and 0.15 mm. that the thermo-electric power when coupled with silver varied from 75 to 82 microvolts per degree, depending upon the purity of the material. Haken (Verh. Phys. Gesell., 1910, 12, p. 229) and Gelhoff and

b. *The size of the wire used in forming the couples.* The chief defects in the older types of thermopiles were their great heat capacity and their consequent lag in reaching a temperature equilibrium. The larger the wire used in making the couple, the greater, of course, will be the heat capacity. In the recent attempts that have been made to improve the linear thermopile a prominent item of change has been the use of finer wires, which not only decreases the heat capacity and lag and increases the radiation sensitivity, but permits more elements to be placed in a given area. The decrease in the size of the wire, however, increases the internal resistance which must of course be taken into account in planning for sensitivity. For example, Coblenz⁸ found in experiments with surface thermopiles that a bismuth wire 0.15 mm. in diameter had sufficient heat capacity to require a half minute to attain thermal equilibrium, while a wire 0.1 mm. in diameter gave satisfactory results. Using this wire in conjunction with one of silver 0.0513 mm. in diameter as a standard of sensitivity, a silver wire of 0.041 mm. in diameter gave a sensitivity of 1.13; one of 0.03 mm. in diameter, a sensitivity of 1.20; and one of 0.021 mm. diameter, a sensitivity of only 1.12. That is, when the wire has reached an optimum fineness, any further decrease in size so increases the internal resistance as to more than

Neumeier (*ibid.*, 1913, 15, p. 876) found that an alloy of bismuth with 9 to 10 per cent antimony gives a thermo-electric power which varies from 77 to 87 microvolts. Coblenz (*Op. cit.*, p. 149) found that an alloy of 5 to 6 per cent of tin gives a thermal E. M. F. of -44 to -45 microvolts per degree; and a thermo-element made of high grade bismuth and this alloy gives a thermo-electric power of 125 to 127 microvolts per degree.

While having 50 to 60 per cent higher thermo-electric power than a bismuth-silver pile, piles made of the bismuth alloy showed only about 10 per cent higher radiation sensitivity. The alloy is so much harder to handle that the same nicety of construction is not possible, also as high a durability is not attained. Since in making the silver-bismuth couple a bead of tin is used in soldering the two wires together, an alloy of bismuth and tin is made at the junction.

A bismuth-iron thermo-element (*op. cit.*, pp. 151-154) was found to give a thermal E. M. F. which was 18 per cent higher than was obtained from bismuth and silver. No increase of radiation sensitivity was obtained, however, because the initial resistance was almost doubled by the use of the iron.

⁸ Coblenz, W. W. Bulletin Bureau of Standards, 1913, 9, pp. 21-22.

compensate for the advantage gained by the lessened heat capacity. Johansen further says⁹ that the radii of the two wires of the thermo-element should be so chosen that the ratio between the heat conductivity and the electrical resistance is the same in both.

c. *The dimensions of the pile and the number and arrangement of the receiving thermo-couples.* In a recent theoretical contribution to the construction of thermopiles for the measurement of radiant energy, more especially the construction of vacuum thermopiles, Johansen¹⁰ arrives at the conclusion that the radiation sensitivity is proportional to the square root of the exposed surface in case of the thermopile as it is in case of the bolometer. In extensive experimental determinations of the point, however, Coblentz¹¹ finds (a) that in single thermo-couples the sensitivity is not proportional to the square root of the area exposed to radiation, but that the area has an optimum value which gives a considerably higher sensitivity than is required compatible with the square root law; and (b) that the highest sensitivity is attained by building up a composite receiver of elements having individual receivers of a size giving the maximum sensitivity.¹² It is obvious, therefore, that sensitivity can be added to the instrument by increasing the total area of the receiving surface and consequently the number of thermo-couples, the individual receivers of which make up the total area; and that the maximum increase can be attained by having each individual receiver of the optimum size. In one of his more recent models of linear thermopiles

⁹ Johansen, E. S. *Ann. der Phys.*, 1910, 33, (4), p. 517.

¹⁰ Johansen, E. S. *Loc. cit.*

¹¹ Coblentz, W. W. *Bulletin of Bureau of Standards*, 1914, 11, p. 142.

¹² From the data obtained in constructing the receiving surface in this way, he concludes that the gain in sensitivity over what is indicated by the square root law amounts probably to as much as 50 per cent.

According to Coblentz the requisite of the optimum size is that it shall absorb radiant energy at a rate which will just compensate for the loss of heat from conduction along the wires. If this size is exceeded, the loss from emission becomes even greater than the loss by conduction along the wires and the two together operate to give less than the maximum difference of temperature attainable between the "hot" and "cold" junctions of the couple. A lag in reaching a thermal equilibrium also results because the heat is drained off from the center of the receiver faster than from the edges by conduction along the wire.

Coblentz¹³ uses, for example, 22 junctions of bismuth and silver mounted in a space 10.5 mm. long. The width of this pile was 5 mm. and its resistance was 10.8 ohms.¹⁴ In the surface thermopile greater sensitivity may of course be attained than in the linear. The surface pile is in effect built up of contiguous linear piles.

d. *The type of connection of the couples.* In the older types of thermopile it was the custom to connect the couples in series. Coblentz¹⁵ has found, however, that it is of advantage to substitute a series-parallel connection. In the series connection one thermo-couple is attached to each of the overlapping receivers on the front of the pile, while in the series-parallel arrangement two couples are soldered to each receiver. The effect of this type of connection is in the first place to reduce the number of overlapping receivers by one-half. This reduces the superfluous metal at the lap and the amount of insulation required, and gives the apparatus a quicker response. And secondly the internal resistance is reduced to one-fourth what it would be if the elements were all connected in series; so that although their E. M. F. is reduced by one-half by the series-parallel arrangement, there is a gain of from 10 to 12 percent. in radiation sensitivity.

e. *The relation of internal to external resistance.* It has been a commonly accepted principle in the construction of thermopiles that the highest sensitivity is attained when the resistance in the thermopile is equal to the resistance of galvanometer and connecting wires. Rayleigh,¹⁶ for example, in his computation of the thermodynamic efficiency of the thermopile has shown that the useful work done externally attains a maximum when the external resistance is equal to the internal resistance. In these computations only the specific resistances and the thermal conductivities were considered. Altenkirch,¹⁷ 1909, however, contends that the

¹³ Coblentz, W. W. Bulletin of Bureau of Standards, 1913, 9, p. 292.

¹⁴ The linear pile that has been used in the work that has been done in this laboratory is of this type with the exception that the receiving surface, designed for spectroscopic work, has a breadth of only 2 mm.

¹⁵ Coblentz, W. W. Bulletin of the Bureau of Standards, 1914, 11, pp. 138-142.

¹⁶ Rayleigh. Phil. Mag., 1885, 20 (5), p. 361.

¹⁷ Altenkirch, E. Phys. Zeit., 1909, 10, p. 560.

external resistance may be two or three times the internal resistance without seriously affecting the maximum efficiency of the thermopile, and Coblenz,¹⁸ 1914, finds that the external resistance may be two or three times the internal resistance without decreasing the sensitivity of the instrument more than 5 to 10 percent.

f. *Nicety of construction.* Coblenz makes the statement that the attainment of a high radiation sensitivity in a thermopile is at the present stage of development of thermopile making mainly a question of nicety of construction, for upon this more than any other point depends the low heat capacity, conductivity and emissivity needed for a sensitive instrument. The following are some of the points that should be taken into account in attaining the most effective relation between capacity, conductivity and emissivity,—the kind of materials, the size and form of the wires used for the couples, the length of wire, the size of the receiving surface, the type of connection of the couples, the relation of size of slit to size of receiving surface, the amount of insulation material, etc.¹⁹ The object to be attained by a low heat capacity, conductivity and emissivity is of course that the energy falling on the receiving surface shall cause a maximum rise of temperature and that there shall be as little lag as is possible in the rise to

¹⁸ Coblenz, W. W., *op. cit.*, p. 175.

¹⁹ Coblenz attributes a great deal of his success in the construction of thermopiles to the use of his electrically heated welding device. (See Bull. Bureau of Standards, 1913, 9, p. 16); to the choice of silver wire which is easily cleaned and annealed; and to his use of pure tin in the process of welding which produces an alloy which is not brittle.

He cites cases to show the effect on sensitivity of deviations from the general method of construction. For example, the central line of receivers of one thermopile was given an additional coat of shellac to cause the individual receivers to adhere, instead of causing the adhesion by merely moistening the insulating layer with alcohol. The instrument was slow in responding to a radiation stimulus and was besides insensitive. This extra shellac was then removed by means of blotting paper wet with alcohol, and the surfaces resmoked. The radiation sensitivity was increased 40 to 50 per cent. In another case a thermopile was made of 0.1 mm. wire pressed flat. This thermopile had a radiation sensitivity 25 to 30 per cent less than the average sensitivity of a number of similar thermopiles made of round wire. The flat wire which presented a greater surface for radiation increased the emissivity and thus lowered the sensitivity of the instrument.

thermal equilibrium. When this is attained the instrument will respond quickly and give its maximum response.

g. Provisions to secure steadiness of response. The main source of unsteadiness of response is exposure to air currents. This of course can be completely eliminated by isolating the instrument from the air. The best success of isolation is evacuation which doubles the sensitivity. Water jackets and combinations of water and air jackets have been used also. Unlike the bolometer, however, the thermopile is noted for its steadiness of response in air. This is one of the strongest recommendations for its general use.

Older forms of thermopiles used by Melloni, Tyndall and others were subject to a "drift"; *i.e.*, there was a permanent E. M. F. which caused a permanent deflection of the galvanometer. This permanent E. M. F. seems at least in part to have been due to lack of symmetry in the construction of the "hot" and "cold" junctions. In our own linear thermopile this tendency to drift was overcome by soldering on the "cold" junctions receiving surfaces of tin of the same dimensions as were carried by the "hot" junctions.

2. *Advantages of the thermopile.* (1) It is non-selective in its response to wave-length. (2) It is readily portable and is easily adapted to the many needs for which a sensitive radiometer is needed. (3) In its most improved forms it is very steady in its action. Even when used in air there is comparatively little drift. (4) A high degree of sensitivity has been obtained. Coblenz²⁰ with a single thermo-couple in a vacuum and a 3-foot telescope has recently made quantitative measurements of the radiations of stars of the fifth magnitude and detectable responses were obtained from stars of the seventh magnitude. The instruments used by us are abundantly sensitive to measure the visible spectrum. (5) It is already attainable in forms adapted to special purposes. Coblenz²¹ for example, describes thermopiles for the following purposes: for stellar

²⁰ Coblenz, W. W. Publications of the Astronomical Society of the Pacific, 1914, 26, pp. 169-178.

²¹ Coblenz, W. W. Bulletin of Bureau of Standards, 1914, 11, p. 163.

measurements, and the measurements of other nocturnal radiations; for the measurements needed in physical photometry; for the determination of whether or not heat is generated in the tetanization of a nerve (an ingenious device in which the thermo-couple is made into a U-shaped trough for the reception of the nerve); and for various miscellaneous purposes which need not be gone into here. Its feasibility and wide range of utility are attested by the fact that it is now being used with success and a fair degree of convenience in radiation work in physical, chemical, biological and psychological laboratories. Owing to the recent improvements that have been made in its sensitivity, its quickness and steadiness of response, and the ease and convenience with which it can be operated, it seems to be the most promising of the radiation instruments now available, especially for the use of the experimenter who is not a radiometric specialist. These improvements mark, it is to be hoped, an epoch in the quantitative study of phenomena in the production of which radiation plays a part.

B. THE NICHOL'S RADIOMETER

The radiometer was first described by Crookes²² in 1874 as an interesting scientific toy. Some years later it was used in a somewhat modified form to investigate the infra-red spectrum to about 1.5μ .²³ The first really useful radiometer was developed by Nichols in 1896.²⁴ It was further developed and improved by Coblentz in 1905.²⁵ In its modern form the radiometer consists of two similar thin vanes of mica or platinum blackened on one side which are held together by glass fibres and are suspended in a vacuum by means of a fine quartz fiber. The vanes are about 3 mm. from an opening or window in the housing of the apparatus. The radiations to be measured fall upon one of the vanes which becomes slightly warmed. This

²² Crookes, W. *Philos Trans.* 1874, 164, p. 501; 1875, 165, p. 519; 1876, 166, p. 325.

²³ Pringsheim, E. *Ann. der Phys.* 1883, 18, p. 32.

²⁴ Nichols, E. F. *Berichte der Berliner Akad.*, 1896, p. 1186; *Phys. Rev.*, 1897, 4, p. 297.

²⁵ Coblentz, W. W. *Investigations of Infra-red Spectra*. Carnegie Publication, No. 35, Washington, 1905, p. 21.

causes the residual gas molecules to rebound with increased velocity from the blackened surface and the reaction pushes this vane from the window causing a rotation about the axis of suspension. A small mirror is attached to the glass fiber which forms the axis of rotation, and the deflection is observed by means of a telescope and scale.

1. *Significant points with regard to the radiometer.* The behavior of the radiometer has been worked out theoretically by Maxwell²⁶ in his paper on "Stresses in Rarefied Gases Arising from Inequalities of Temperature." Crookes, Nichols, and others have shown that the sensitiveness of the radiometer is a function of the pressure of the residual gas surrounding the vanes, of the kind of gas, and of the distance of the exposed vanes from the window. Investigation by Coblentz²⁷ has also brought out the following points. (1) For vanes of small dimensions such as must be used in practical work, the deflections are found to be proportional to the area of the exposed surface of the vane. (2) The sensitiveness varies with the diameter of the suspension fiber. (3) The instrument is not selective in its response.

2. *Comparative advantages and disadvantages of the radiometer.* As a working instrument the radiometer is said to have the following advantages. (a) Its sensitiveness is comparatively easy to control since it can be made to depend almost entirely upon the pressure of the residual gas. (b) It is not influenced by magnetic and thermo-electric disturbances which render work with a very sensitive galvanometer tedious and unsatisfactory. (c) It is not so sensitive to temperature changes as is, for example, a bolometer, and it can be more easily shielded from changes in temperature than can a bolometer with its galvanometer, battery, etc. It has the following disadvantages. (a) It is not portable, which may cause inconvenience in certain types of work. (b) For maximum sensitiveness its period is very long as compared with that of a bolometer or thermopile and galvanometer combination. (Nichols used a period of 8-12 sec., single

²⁶ Maxwell, J. C. The Scientific Papers of, 2, p. 681; Philos. Trans., 1879, 170, p. 231.

²⁷ Coblentz, W. W. Bull. Bur. Standards, 1907, 4, p. 405.

swing; Coblentz 30-45 sec.) This makes the instrument slow to operate. As a compensating feature, however, as Coblentz points out, the readings are always trustworthy so that there is no need to repeat them. (c) Its window or preferably double window is selective in its transmission in the invisible parts of the spectrum. A correction has, therefore, to be applied to the results for this inequality when working in this region.

C. THE RADIO-MICROMETER.

This instrument was invented independently by d'Arsonval²⁸ and by Boys.²⁹ It combines in one instrument the thermocouple which in response to the radiant energy generates the electric current, and the galvanometer which indicates by its deflections the comparative amounts of current. That is, the radio-micrometer is essentially a moving coil galvanometer, the moving coil of which contains one or more thermojunctions. In the instrument devised by d'Arsonval a single loop of wire was used, one part of which was silver and the other palladium. In the instrument devised by Boys the moving coil consisted of a loop of copper wire to which was soldered a thermo-junction of bismuth and antimony. These instruments not having been found to possess the sensitivity attributed to them by their inventors, various attempts have been made to improve them but with little success. Paschen,³⁰ for example, tried to increase the sensitivity by increasing the number of thermojunctions.³¹ Different thermo-couples have been employed.

²⁸ d'Arsonval. *Soc. Franc. de Phys.*, 1886, pp. 30, 77.

²⁹ Boys, C. V. *Proc. Roy. Soc.* 1887, 42, p. 189; 1888, 44, p. 96; 1890, 47, p. 480; *Philos. Trans.*, 1889, 180A, p. 159.

³⁰ Paschen, F. *Ann. der Phys.*, 1893, (3) 48, p. 272.

³¹ An advantage is gained in the thermopile by increasing the number of thermo-couples, but not in the radio-micrometer. In a thermopile the highest efficiency is attained when the resistance of the thermo-couples is equal to the combined resistance of the connecting wires and the auxiliary galvanometer. Since the resistance of a single thermo-couple is much less than this combined resistance, it is of advantage to use several pairs of junctions. In the case of the radio-micrometer, however, the connecting loop of wire has a negligible resistance, hence there is nothing to gain by using more than a single pair of junctions; for as the electromotive force is increased by the addition of junctions, there is a proportionate increase of resistance and the throw of current remains constant.

Féry,³² for example, used silver and constantan; Schmidt³³ bismuth and antimony; and Coblentz,³⁴ also bismuth and antimony, and later bismuth and silver.³⁵ Hollnagel³⁶ added greatly to the sensitivity and constancy of the instrument described by Schmidt by operating it in a vacuum; Coblentz³⁷ increased the sensitivity of his bismuth-silver radio-micrometer by enclosing it in a vacuum; and Rubens and Hollnagel,³⁸ and Rubens and Wood³⁹ succeeded in obtaining an increase of sensitivity for the instrument described by Schmidt by using a concentrating or conical receiver. Coblentz also found that the sensitivity of his instrument was lowered by para- and dia-magnetic effects produced by the field magnets. He was able to lessen these effects and add thereby to the delicacy of response by using weak field magnets; or if strong, by placing them as far above the thermo-junctions as was possible. The elimination of these effects he considers one of the chief obstacles to be overcome in the future construction of the instrument.⁴⁰

As a working instrument the radio-micrometer may be said to have the following advantages: (a) It is self-contained; (b) it is non-selective in its response to wave-length; (c) it is little subject to magnetic perturbations; and (d) it has a high constancy of the zero-reading.

³² Fery, C. *Comptes Rendus*, 1909, 148, p. 915.

³³ Schmidt, H. *Inaug. Diss.*, Berlin, 1909; *Ann. der Phys.*, 1909, 29 (5), p. 1004. See also U. Meyer. *Ann. der Phys.*, 1909, 30 (5), p. 612.

³⁴ Coblentz, W. W. *Bulletin Bureau of Standards*, 1906, 2, p. 479.

³⁵ On p. 10 (*Bulletin Bureau of Standards*, 1913, 9), Coblentz says: "From later experience it seems desirable to try constantan instead of bismuth."

³⁶ Hollnagel, H. *Inaug. Diss.*, Berlin, 1910.

³⁷ Coblentz, W. W. *Bull. Bureau of Standards*, 1906, 2, p. 479; 1908, 4, p. 396.

³⁸ Rubens, H. and Hollnagel, H. *Sitz Ber. d. könig. Preuss. Akad. Wiss.*, Berlin, 1910, No. 2, p. 26.

³⁹ Rubens, H. and Wood, R. W. *Ibid.*, 1910, No. 52, p. 1122.

⁴⁰ For further reports of work with the radio-micrometer see Lewis, E. P. *Astrophysical Journal*, 1895, 2, p. 1; Wilson, W. E. *Proc. Roy. Soc.*, 1894, 55, p. 246; 1895, 58, p. 174; 1896, 60, p. 377; and Julius, W. T. *Handlungen*, 5, de *Nederlandisch Natuur en Geneeskundig Congress*, 1895.

D. THE BOLOMETER.

As has already been stated the bolometer is an instrument depending for its response to radiant energy on the change in resistance with change in temperature offered by a metal to the flow of an electric current. The instrument is essentially a Wheatstone bridge, two arms of which are made of very thin blackened metal strips (of high electrical resistance and high temperature coefficient), one or both of which are exposed to radiation. When thus exposed there is a change of temperature which unbalances the bridge, and the resulting deflection of the needle of the galvanometer in circuit with the bridge gives a measure of the energy absorbed. In order that the instrument shall be sensitive to small radiation quantities, it is obvious that the metal used should have a high temperature coefficient of resistance, a small specific heat, and a low heat conductivity. Such metals are nickel, platinum, tin, and iron. For various reasons relating to mechanical construction, however, platinum is much more frequently used than the others.

The earliest account of an instrument depending on change of electrical resistance for measuring or detecting radiant energy appears to be that of Svanberg, 1851,⁴¹ who for this purpose introduced a flat spiral of blackened copper wire into one of the arms of a Wheatstone bridge. Langley, 1881,⁴² was the first however, to invent a practical instrument and demonstrate its superiority to all radiation meters existing at that time for accuracy, quickness of action, and adaptability. As is shown in the reference appended below, his improvements of the instrument extended over a long period of time. The value of these improvements may be shown by comparing the sensitivity of his earlier and later instruments. The first had a sensitivity of 0.00002° per mm. deflection of the galvanometer, and the latest recorded a temperature change of 0.000001° per mm. deflection when used with a galvanometer having a figure of merit of $i=5 \times 10^{-10}$ ampere.

⁴¹ Svanberg, A. F. *Pogg. Ann. der Phys.*, 1851, 84, p. 416.

⁴² Langley, S. P. *Proc. Amer. Acad.*, 1881, 16, p. 342; *Chemical News*, 1881, 43, p. 6; *Brit. Assoc. Report*, 1894, p. 465; *Annals Astrophys.*, Obs. I.

1. *Important points in the construction of sensitive bolometers.* Obviously effective sensitivity can be added in two ways in the use of the bolometer: (1) by improving the bolometer itself; and (2) by making more delicate the auxiliary galvanometer. In the attempt to construct sensitive bolometers with as great constancy of the zero as is possible, the following are some of the points that have received attention.

a. *The kind of material to be used for the receiving surface.* As has been stated the problem is to get a metal having a high temperature coefficient of resistance, low specific heat; and low conductivity of heat.⁴³ The following metals have at various times been used: platinum,⁴⁴ tin,⁴⁵ nickel,⁴⁶ and iron.⁴⁷

b. *The area of the receiving surface.* The sensitiveness has been found to be closely proportional to the square root of this surface. In spectral energy work, therefore, where the bolometer strip is narrow the sensitiveness attainable for the bolometer is limited.

c. *The thickness of the strip used as receiver.* There is a mechanical limit to the thinness of the strip that can be used when exposed to the air. Langley⁴⁸ found that platinum strips, for

⁴³ According to Lummer and Kurlbaum (Wied. Ann. der Phys., 1892, 46, p. 208) the following equation expresses the relation between the sensitiveness of the bolometer, S ; the bolometer current, I ; the temperature coefficient of the area exposed to radiation, e ; the area of the strip exposed to radiation, a ; the resistance of the bolometer strips, r ; the absorption coefficient of the surface exposed to radiation, A ; the emissivity of the whole surface, E ; the area of the whole surface, F ; the heat capacity, W ; and the galvanometer constant, k .

$$S = \frac{I}{8} e a k \sqrt{r \frac{f_1(A) f_2(F)}{f_2(E) f_1(W)}}$$

From this equation it will be seen that the sensitiveness is increased by decreasing the heat capacity and the emissivity; and by increasing the bolometer current, the temperature coefficient, the resistance, the absorption coefficient, and the surface.

⁴⁴ Langley, S. P. *Loc. cit.*

⁴⁵ Angström, K. Wied. Ann. der Phys., 1885, 26, p. 253; 1889, 36, p. 715; 1893, 68, p. 493, and others.

⁴⁶ Julius, W. T. Licht und Wärmestrahlung, 1890, p. 31.

⁴⁷ Rubens, H., (used tin, iron, and platinum). Wied. Ann. der Phys., 1889, 37, p. 255; and 1892, 45, p. 238.

⁴⁸ Langley, S. P. *Op. cit.*

example, less than 0.002 mm. thick are inadvisable, thinner ones being disturbed mechanically by air currents. In recent work Coblentz⁴⁹ finds it permissible to use platinum strips 0.001 mm. in thickness. In vacuum bolometers, however, much thinner strips (0.0005 mm.) are used to advantage.

d. *The most favorable resistance of bolometer and balancing coils.* Lummer and Kurlbaum⁵⁰ considered the bolometer a simple Wheatstone bridge which has its maximum sensitiveness when the four arms and the galvanometer are all of equal resistance. In fact in the construction of their instrument instead of using one or two bolometer surfaces, they used four just alike, each forming one of the arms of the bridge.⁵¹ Child and Stewart,⁵² however, have shown experimentally that the sensitiveness is increased by having the resistance of the balancing coils several times that of the bolometer strips. Abbot⁵³ has also shown that the maximum sensibility is very nearly attained when the resistance of the balancing coils is four times or more that of the bolometer strips, and the galvanometer resistance is not less than 0.6 or more than four times the resistance of the bolometer strips.⁵⁴

e. *The slide wire for balancing the resistance.* The question of a satisfactory slide wire has required considerable attention. Coblentz has found that slide wires of platinum 0.5 and 1 mm. in diameter used in connection with a mercury contact give the best satisfaction.

f. *The protection of the bolometer from air currents.* This can be done adequately only by putting the bolometer in a vacuum. Contact with the air renders the bolometer both insensitive and inconstant in its response. On the former point,

⁴⁹ Coblentz, W. W. Bulletin of the Bureau of Standards, 1912, 9, p. 37.

⁵⁰ Lummer, O. and Kurlbaum, F. Wied. Ann. der Phys., 1892, 46, p. 204.

⁵¹ The tendency among foreign investigators has been to make the four arms of equal resistance. On the other hand just as strong a tendency has been shown among American investigators to make the resistance of the balancing coils greater than that of the bolometer strips.

⁵² Child, C., and Stewart, O. Phys. Rev., 1897, 4, p. 502.

⁵³ Abbot, C. G. Annals of the Astrophysics Obs., I.

⁵⁴ See also Reid, H. F. Amer. Jour. Sci., 1888, 33, (3), p. 160.

Warburg, Leithäuser, and Johansen⁵⁵ have shown that the heat lost by air conduction for a bolometer 1 mm. wide is 4.5, and for a bolometer 0.2 mm. wide, 14.8 times as great as it is from radiation. And in a vacuum a bolometer 0.2 mm. wide when operated with a small current was found to be ten times as sensitive as it was in air.⁵⁶

g. *The strength of current.* The radiation sensitivity of a vacuum bolometer is found to be proportional to the current for small values but for a large current the radiation sensitivity of a narrow bolometer passes through a maximum. This maximum is obtained for a current density at which the radiation sensitivity of the air bolometer does not depart appreciably from proportionality with the current. The manner in which the radiation sensitivity varies with the gas pressure and with the bolometer current is shown, for example, by Buchwald.⁵⁷

2. *Points in the construction of the auxiliary galvanometer.* It is scarcely necessary to mention that the effective sensitivity of the bolometer depends in a large measure upon the auxiliary galvanometer. The first great step in improving the moving magnet galvanometer is due to Kelvin who decreased the weight of the moving parts to a few milligrams, and introduced the static system of magnets. Snow⁵⁸ was among the first to give much attention to the possibility of adding sensitivity to the bolometer by improving the galvanometer. Paschen⁵⁹ continued the work in this direction and constructed the most sensitive galvanometer used up to that time. DuBois and Rubens,⁶⁰ Mendenhall and

⁵⁵ Warburg, E., Leithäuser, G. and Johansen, E. *Ann. der Phys.*, 1907, 24, (5), p. 25.

⁵⁶ A noteworthy vacuum spectro-bolometer is described by A. Trowbridge (*Phys. Rev.*, 1908, 27, p. 282; *Philos. Mag.*, 1910, (6), 20, p. 768) in which the bolometer and the optical parts of the spectroscope are in a vacuum. Coblenz also describes a very sensitive vacuum bolometer and gives results with it at different pressures (*Bull. Bureau of Standards*, 1913, 9, pp. 39-43). Other less adequate methods of shielding have been to surround the bolometer by a double wall with an air space between; to enclose it in a water jacket (Langley, *op. cit.* and Abbot, *op. cit.*); etc.

⁵⁷ Buchwald, E. *Ann. der Phys.*, 1910, (4), 35, p. 928.

⁵⁸ Snow, B. W. *Phys. Rev.*, 1895, 1, p. 31.

⁵⁹ Paschen, F. *Wied. Ann. der Phys.*, 1893, 48, p. 272.

⁶⁰ DuBois and Rubens. *Ann. d. Phys.*, 1900, (4), 2, p. 84.

Waidner,⁶¹ Abbot,⁶² Ingersoll,⁶³ and Coblentz⁶⁴ have all described sensitive galvanometers. Some of the points to be considered in the construction of a sensitive galvanometer are form and size of coil, size of wire, the kind of magnet and the dimensions and construction of the needle system, the astaticizing of the magnet system, the shielding of the system from influences due to the earth's field and neighboring objects, etc. The proper form and method of winding galvanometer coils to secure a maximum effect from a given weight or resistance of copper has been thoroughly discussed by Maxwell.⁶⁵ He shows that the greatest effect is obtained by winding the coils with different sizes of wire, beginning with the smallest size and winding each layer so that it lies within the surface the polar equation of which is $r^2 = d^2 \sin \theta$, where r is the length of the radius making an angle θ with the axis of the coil, and d the value of r when $\theta = 90^\circ$. Abbot⁶⁶ has computed the most efficient coils for meeting these conditions, and gives results for coils wound with a single wire and for coils wound with three sections of wire of different diameters. He found that the total force exerted at the center is closely proportional to the 0.45 power of the total resistance and that coils composed of three sections of the best sizes of wire give 1.4 times the force of a coil of the best single size wire of the same total resistance. In his best 25 ohm coil, wound in three sections the diameters of the wires were 0.08, 0.16, and 0.32 mm.; the lengths were 256, 1031, and 4144 cm.; and the external diameter of the completed coil was 3.3 cm.

In the construction of the needle system the greatest sensitivity is attained when the ratio of the magnetic moment to the moment of inertia of the system is a maximum. The best dimensions and construction of needle systems have been extensively investigated

⁶¹ Mendenhall, C. E. and Waidner, C. W. *Amer. Journ. Sci.*, 1901, 12, (4), p. 249.

⁶² Abbot, C. G. *Astrophys. Journ.*, 1903, 18, p. 1.

⁶³ Ingersoll, L. R. *Philos. Mag.*, 1906, (6), 11, p. 41.

⁶⁴ Coblentz, W. W. *Bulletin Bureau of Standards*, 1908, (3), 4, pp. 424-435; 1916, 13, p. 423.

⁶⁵ Maxwell, J. C. *Electricity and Magnetism*, 2, p. 322.

⁶⁶ Abbot, C. G. *Loc. cit.*

by Paschen, Mendenhall and Waidner, and by Abbot. The shielding of the galvanometer from magnetic perturbations, etc., is done by means of housings of soft iron. For an inexpensive and convenient method of shielding, for a simplification of the moving coil galvanometer for convenience of shielding, and for the astaticization of the needle system, see Coblentz (*loc. cit.*).

3. *Possible sources of difficulty in the use of the bolometer.* The preceding discussion though brief may be enough to indicate that the bolometer is a difficult instrument to operate. The following are the possible sources of trouble. (1) The auxiliary galvanometer is subject to magnetic perturbations and if exposed to great temperature changes its sensitiveness is changed, due to a variation in the resistance of the coils. The sensitiveness and zero reading are also subject to frequent changes due to variations in the magnetic field. (2) The bolometer strip is affected by air drafts, and, if very thin, by mechanical vibration. (3) The electric circuits are subject to temperature (resistance) changes which cause variations in the bolometer current. (4) The storage battery current is irregular due to changes in temperature and to polarization. (5) The gases surrounding the bolometer may affect the readings. Lummer and Pringsheim⁶⁷ found, for example, that variations in the amount of moisture in the air change the sensitiveness of the bolometer. A part or all of these causes tend to make the readings in work with the bolometer extremely variable. These variations are of two kinds. (a) A slow drift of the zero scale reading due to changes in the resistance of the bridge; and (b) fluctuations of the reading due to air currents and magnetic perturbations.

4. *The comparative advantages and disadvantages of the bolometer.* The bolometer has, however, the following advantages. It has a high degree of sensitivity. It is portable. It is non-selective in its response to wave-length. It can be calibrated directly against a black body. It is quick in its action and is, therefore, well adapted to work in which a quick registration of the galvanometer deflections is desired. Its chief disadvan-

⁶⁷ Lummer, O. and Pringsheim, E. *Ann. der. Phys.*, 1897, (3), 63, p. 398.

tage, as has been stated, is its unsteadiness and the difficulty of operating. It is not nearly so easy to operate as the thermopile, for example; and the unsteadiness of its zero point renders it untrustworthy for small readings in spite of its high intrinsic sensitivity.

E. THE SELENIUM CELL

To Willoughby Smith belongs the discovery that has led to the use of selenium as a light-measuring instrument. In 1873 while using a resistance made of selenium in connection with experiments in telegraphy, he discovered that its electrical conductivity is raised by exposure to light. The immediate result of this discovery was twofold: light-measuring instruments were constructed of selenium, and a long series of investigations was begun to determine (a) the factors extraneous to light that influence the resistance of selenium and thus affect its applicability to the measurement of light; (b) the factors that influence the action of light on the conductivity of a selenium cell and the possibility of the use of the cell either as a photometer or a radiometer; and (c) the nature of the action of light on the specific resistance of selenium.

A selenium cell is a device consisting essentially of a mass of crystalline selenium furnished with two metallic electrodes. Crystalline selenium is obtained by keeping molten vitreous selenium at a temperature of from 150 to 210° C. for several hours. It then takes on a metallic appearance and becomes opaque even in case of very thin films. Selenium has such a very small conductivity that in making resistances of it, one feature of the construction is to offer several paths or rather one continuous broad path for the flow of current. One way in which this has been accomplished in the construction of light-measuring cells has been to wind a strip of mica or slate with two parallel wires less than 1 mm. apart. This is covered with powdered selenium. The selenium is melted, worked into a smooth surface and cooled quickly. It is then heated again and cooled very slowly. One end of each wire is connected to the battery terminals, the others end in the selenium. The wires thus in reality form the electrodes, and the circuit is completed through the intervening selenium. This type

of cell has been used among others by Bidwell, Siemans, Sabine, and Pfund. (For construction of selenium cells, see Bidwell,⁶⁸ Fritts,⁶⁹ Berndt,⁷⁰ Townsend⁷¹ and Dieterich.⁷²)

1. *Points to be considered in the construction and use of selenium cells.* The following are some of the points to be considered in the construction and use of sensitive selenium cells.

a. *The method of preparation.* The sensitiveness of the cell to light depends largely upon its initial specific resistance. This has been pointed out by Pochettino⁷³, Giltay,⁷⁴ and especially by Brown.⁷⁵ According to Brown, the higher the resistance of crystalline selenium the greater is its sensitivity to light; and conversely the lower its resistance, the less is the sensitivity. Brown gives results showing the resistance of the cell and its sensitivity in terms of the ratio of conductivity in light to conductivity in the dark. For example, a cell with a resistance of 10^9 ohms had a sensitivity of 200:1 in an arbitrary scale; a cell with a resistance of 400,000 ohms, a sensitivity of 30:1; 3,000 ohms, 2:1; 1,700 ohms, 1:1. In its vitreous state selenium is practically a non-conductor. To become a conductor it must be brought to the crystalline form. For example, when it is heated to a temperature of 100 to 150° C. its conductivity is slight and variable; but when heated repeatedly to temperatures of 190 to 210° C. and cooled, it passes into a coarsely granular crystalline state and acquires and retains a greater conductivity. That the temperature to which selenium has been heated is the chief factor in determining its conductivity and sensitivity to light was pointed out by Siemans⁷⁶ as early as 1875, who stated that when heated to 210° C. cells of greater conductivity, constancy, and light sensitivity were produced than when heated to 150° C. A systematic study of the

⁶⁸ Bidwell, S. Phil. Mag., 1891, Ser. 5, 31, pp. 250-256; 1895, 40, pp. 233-256.

⁶⁹ Fritts. Electrical Review, 1885, p. 208.

⁷⁰ Berndt, G. Phys. Zeit., 1904, 5, pp. 121-124.

⁷¹ Townsend, F. Electrician, Oct. 7, 1904, 53, pp. 987-990.

⁷² Dieterich, E. O. Phys. Rev., 1914, 4, Ser. 2, pp. 467-476.

⁷³ Pochettino, A. N. Cimento, 1911, 1, Ser. 6, pp. 147-210.

⁷⁴ Giltay, J. W. Phys. Zeit., 1910, 11, p. 419.

⁷⁵ Brown, F. C. Phys. Rev., 1911, 33, pp. 1-26.

⁷⁶ Siemans, W. Phil. Mag., 1875, 50, p. 416.

effect of temperature and the duration of annealing on the resistance of selenium has been made by Dieterich.⁷⁷ His results show that when maintained at a temperature of 200 to 210° C. for six hours, a cell was produced with a resistance of 233,000 ohms; at 210° for four hours, a resistance of 358,000 ohms; at 210° for five hours, 490,000 ohms; at 180° for three and one-half hours, 1,400,000 ohms; and at 190° for two hours, 3,690,000 ohms. Inasmuch as his cells of highest resistance were not permanent, he was not able unfortunately to work out the correlation between resistance and sensitivity. Pochettino,⁷⁸ Aichi and Tanakadate,⁷⁹ Brown,⁸⁰ and Dieterich⁸¹ all think that a change of structure takes place when selenium is annealed at a temperature of 210 to 220° which causes the increase of conductivity.

b. *Purity of the selenium.* Bidwell⁸² found that insensitive selenium cells were increased in sensitivity by the addition of a small quantity of cuprous selenide. Marc⁸³ recommends the addition of 0.1-0.5% of silver to increase its sensitivity. Townsend⁸⁴ claims that 1 or 2% of copper or nickel selenide may be present without affecting to a marked degree the sensitivity of the cell. Pfund⁸⁵ found that the sensitivity could be increased by the presence of 3% of a selenide. He believes it to be of advantage, however, to start with a chemically pure selenium and add impurity of a definite kind and amount.

c. *Material and size of electrodes.* Sale,⁸⁶ and Adams and Day⁸⁷ used platinum electrodes; Bell⁸⁸ used brass; Bidwell,⁸⁹

⁷⁷ Dieterich, E. O. *Loc. cit.*

⁷⁸ Pochettino, A. N. *Cimento*, 1911, 1, Ser. 6, pp. 147-210.

⁷⁹ Aichi, K. and Tanakadate, T. *Math. and Phys. Soc., Tokyo*, 1904, (2), 16, pp. 217-221.

⁸⁰ Brown, F. C. *Loc. cit.*

⁸¹ Dieterich, E. *Loc. cit.*

⁸² Bidwell, S. *Phil. Mag.*, 1895, 40, Ser. 5, pp. 233-256.

⁸³ Marc, R. Z. *Anorg. Chem.*, 1906, 48, pp. 393-426.

⁸⁴ Townsend, F. *Loc. cit.*

⁸⁵ Pfund, A. H. *Phil. Mag.*, 1904, 7, Ser. 6, pp. 26-39.

⁸⁶ Sale. *Proc. Roy. Soc. of London*, 1873, 21, pp. 283-285.

⁸⁷ Adams and Day. *Philos. Trans.*, 1877, 167, pp. 313-349.

⁸⁸ Bell, G. *Nature*, 1878, 22, p. 500.

⁸⁹ Bidwell, S. *Loc. cit.*

copper; Pfund⁹⁰ and Berndt,⁹¹ carbon. Dieterich⁹² tried copper, nickel, platinum, German silver and Advance wire. He found that copper, German silver and Advance wire have the disadvantage that at the temperature of annealing a film of oxide is formed. This so materially increases the resistance of the cell as to make it practically useless except with very sensitive auxiliary apparatus. Nickel wire is much less easily oxidized and proved as satisfactory as platinum wire besides being less expensive. Pfund and Berndt consider carbon electrodes preferable in that selenium forms no conducting compound with carbon. On the question of size of electrodes, there seems to be general agreement that large surface contact between the junctions and the selenium is necessary to avoid high junction resistance and consequent diminished sensitivity of the cell.

d. *Strength of the battery current.* The sensitivity of the selenium cell has been found to vary with the battery current. It was first noted by Adams and Day⁹³ that the resistance of selenium diminished as the battery current is increased. Sabine⁹⁴ held this to be true only after a certain intensity of current had been reached. For lower intensities of current, increase of current caused increase of resistance. Minchin⁹⁵ increased the conductivity of his cell fourfold by increasing the voltage from 2 to 12 volts. Brown⁹⁶ found that with a Ruhmer cell the conductivity varied by an amount almost directly proportional to the voltage; with a Giltay cell, however, the variation decreased in amount as the voltage was increased. Ries⁹⁷ claims that conductivity increases with increase of voltage for a range of from 0.4 to 4 volts. Luterbacher⁹⁸ states that this change is greater for direct than for alternating current. The necessity for an accurately

⁹⁰ Pfund, A. H. *Loc. cit.*

⁹¹ Berndt, G. *Loc. cit.*

⁹² Dieterich, E. O. *Phys. Rev.*, 1914, 4, Ser. 2, p. 468.

⁹³ Adams and Day. *Op. cit.*, p. 319.

⁹⁴ Sabine, R. *Phil. Mag.*, 1878, 5, ser. 4, pp. 401-416.

⁹⁵ Minchin, G. M. *Phil. Mag.*, 1891, 31, ser. 5, pp. 207-238.

⁹⁶ Brown, F. C. *Loc. cit.*

⁹⁷ Ries, C. *Phys. Zeit.*, 1911, 12, p. 529.

⁹⁸ Luterbacher, J. *Ann. der Physik*, 1910, 33, p. 1392.

constant battery current when using a selenium cell as a measuring instrument is obvious.

e. *The direction of the battery current.* Adams and Day⁹⁹ found that the passage of a current in any direction at any period in a series of observations produces a condition which tends to facilitate the subsequent passage of a current in the opposite direction but obstructs one passing in the same direction. He interprets this condition as a slight "set" of the molecules. The effect is particularly marked in case of the first current sent through the selenium and is more or less permanent. This result was confirmed by Sabine¹⁰⁰ who thinks the changes are in the resistance of both the selenium and the junctions. This fact combined with the changes in resistance caused by changes in the strength and duration of current led Sabine to state that selenium is very unsuitable for the production of a constant resistance for measuring purposes.

f. *Duration of battery current.* Adams and Day¹⁰¹ found that the resistance of selenium increases continuously during the time of the passage of the battery current. They point out, for example, that on this account the precaution should always be taken to shut off the current between observations. This precaution, however, does not eliminate; it only lessens the effect of the variable factor.

g. *Temperature.* Bidwell¹⁰² claims that there is an optimum temperature for each cell above and below which the resistance decreases. For six cells this temperature was 24, 23, 14, 30, 25 and 22° C. Brown and Stabbins¹⁰³ tested the effect of temperatures ranging from 40° to 200° C. and found that the resistance of selenium decreases with increase of temperature. Temperatures above or below these were not used, so their results contain nothing that bears directly on the claim made by Bidwell. For a change of temperature ranging from 13.2° to 73.4° C. they found that a given amount of light incident on the cell caused changes of

⁹⁹ Adams and Day. *Op. cit.*, p. 323.

¹⁰⁰ Sabine, R. *Loc. cit.*

¹⁰¹ Adams and Day. *Op. cit.*, p. 314.

¹⁰² Bidwell, S. *Phil. Mag.*, 1881, *II*, Ser. 5, p. 302; 1895, *40*, ser. 5, p. 242.

¹⁰³ Brown and Stebbins. *Phys. Rev.*, 1908, *26*, pp. 273-298.

resistance varying in percentage from 24.9 to 3.7. The effect of temperature on the sensitivity of the cell is so marked that Pfund, for example, worked in a room in which the temperature was kept constant to $1/10^{\circ}$.

h. *Pressure*. According to Brown,¹⁰⁴ Brown and Stebbins,¹⁰⁵ and Montén,¹⁰⁶ increase of pressure decreases the resistance of selenium and lowers its sensitivity to light. Brown found that these effects were present up to a pressure of 1,000 atmospheres. In case of a single crystal of selenium, he increased the conductivity about 120 times by an increase of pressure of 180 atmospheres. Brown and Stebbins found the percentage change of resistance for one atmosphere to vary between 0.05 and 0.30 for different cells.

i. *Moisture*. Ries,¹⁰⁷ Bidwell¹⁰⁸ and others have shown that humidity affects the electrical properties of selenium. Ries thinks this effect is sufficient to explain the discrepancies existing in the results of different observers. On this account cells of the Giltay type, which are constructed so that there is free communication between the outer air and the selenium surface, show wide variations in conductivity. Nicholson¹⁰⁹ improved the constancy of a cell of this type by enclosing it in an air-tight box with a glass window.

j. *Age of cell*. Adams and Day¹¹⁰ found the sensitivity of the selenium cell to be greatly reduced after one year. Bidwell¹¹¹ found no material loss at the end of one year, but the cells were practically useless after four years. Dieterich,¹¹² however, constructed two cells of remarkably high sensitivity which lost $\frac{4}{5}$ of their sensitivity within a month.

¹⁰⁴ Brown, F. C. Phys. Rev., 1905, 20, pp. 185-186; Phys. Rev., 1914, 4, ser. 2, pp. 85-98.

¹⁰⁵ Brown and Stebbins. *Loc. cit.*

¹⁰⁶ Montén, F. Ark. för. Mat. Astron. och. Fysik, Stockholm, 1908, 4, pp. 1-6.

¹⁰⁷ Ries, C. Phys. Z., 1908, 9, pp. 569-582.

¹⁰⁸ Bidwell, S. Phil. Mag., 1895, 40, ser. 5, p. 245.

¹⁰⁹ Nicholson, P. J. Phys. Rev., 1914, 3, Ser. 2, p. 8.

¹¹⁰ Adams and Day. *Op. cit.*, p. 348.

¹¹¹ Bidwell, S. Phil. Mag., 1891, 31, Ser. 5, pp. 250-256.

¹¹² Dieterich, E. O. Phys. Rev., 1914, 4, Ser. 2, p. 471.

k. *The amount of polarization gradually set up in the cell.* The presence of polarization currents produced by the passage of a battery current through selenium was found by Adams and Day.¹¹³ This effect was increased by the exposure of the selenium to light. Bidwell¹¹⁴ says the polarization current is very troublesome in making accurate resistance tests by the bridge method. The intensity of this current is increased by humidity. While this factor and the next to be considered, the presence of photo-electric currents, can hardly be said to influence the sensitivity of the cell in a way similar to the preceding factors, they undoubtedly affect its use as a light-measuring instrument; for with the presence of polarization and photo-electric currents of unknown intensity, an exact determination of the conductivity of the cell under a given set of conditions can not be made.

1. *Photo-electric currents.* The presence of photo-electric currents in selenium due to an exposure to light was noted first by Adams and Day,¹¹⁵ later by Bidwell¹¹⁶ and by Minchin.¹¹⁷ Adams and Day found that this current was often more intense than the polarization current and was sufficient to overbalance a weak battery current.

2. *Factors which render it difficult to use the selenium cell for quantitative work either as an ohmic resistance or as a light-measuring instrument.*

A part of the foregoing factors are of importance chiefly in making it almost impossible to construct two selenium cells of similar properties. They do not affect the use of a given cell once constructed. The remainder, however, apply to the responses of a single cell and are so difficult if not impossible of control as to make it exceedingly doubtful whether the selenium cell can be used as an instrument of precision. In fact the consensus of opinion among the investigators has been that it can not be used with a degree of precision which is acceptable in quantitative

¹¹³ Adams and Day. *Op. cit.*, p. 328.

¹¹⁴ Bidwell, S. *Phil. Mag.*, 1895, 40, Ser. 5, p. 244.

¹¹⁵ Adams and Day. *Op. cit.*, p. 333.

¹¹⁶ Bidwell, S. *Op. cit.*, p. 251.

¹¹⁷ Minchin, G. M. *Phil. Mag.*, 1891, 31, Ser. 5, pp. 207-238.

work. These factors are: (1) The passage of the battery current through the cell in a given direction produces a condition which tends to facilitate the subsequent passage in the opposite direction, but obstructs one in the same direction. Since this effect can not be completely eliminated and the cell restored to its original condition by reversing the current, the cell continually changes its state of conductivity with use; hence two measurements can never be made with it in the same condition. This difficulty is further increased by the fact that the longer the current is allowed to flow, the greater is the change of conductivity. The greater number of times the cell is used, therefore, and the longer the current is allowed to flow, the greater will be the progressive change in the properties of the cell. (2) Over and above the effect of current is a loss of sensitivity with age. Measurements made by the cell at intervals at all widely separated are, therefore, not comparable. (3) The polarization currents due to the passage of the battery current and increased by the exposure to light, and the photo-electric currents which are even stronger than the polarization currents and strong enough according to Adams and Day to overcome a weak battery current, produce a variability in the action of the cell for which there seems to be no remedy. When to these apparently insuperable obstacles is added the fact that the strength of the battery current, the temperature of the cell and the humidity of the atmosphere must be kept constant within small limits, one gets some idea of the difficulties attendant on the use of the selenium cell as an instrument of precision.

3. *Factors which apply especially to its use as a light measuring instrument.* The foregoing properties of selenium, it may be noted, apply to its use as material for the construction both of ohmic resistances and of instruments for the measurement of light. In addition to these the following points which apply specifically to its use in the measurement of light are to be considered.

a. *The preexposure of the cell.* Adams and Day¹¹⁸ claim that

¹¹⁸ Adams and Day. *Op. cit.*, p. 315.

selenium is more sensitive in its response to light after it has been kept in the dark for several hours than after it has been exposed to light several times; hence the result obtained from the first of a series of measurements is generally not comparable with those gotten later. Townsend¹¹⁹ says that after prolonged exposure to light there is a fatigue effect which takes place immediately and lasts at least four hours. Nicholson¹²⁰ says that fatigue effects are present when the cell has not been allowed sufficient rest between readings. Marc¹²¹ finds that the sensitivity to red light is greatly modified by a previous strong illumination with white light or by a long continued blue illumination. Grant-ham¹²² investigating the recovery period of the cell, found that for a short time after the exposure to light was cut off, the resistance decreased still further; it then increased rapidly at first, then more slowly. After constant use the cells became at times temporarily almost insensitive.

b. *The time of exposure to light.* For Pfund¹²³ a maximum response was reached in 2 to 3 sec.; then a slight "creeping effect" took place. In later work he¹²⁴ used an exposure time of 12.5 seconds. Brown¹²⁵ claims that the change of conductivity is a function of the time of illumination. For effect of exposure time on response to monochromatic light, see pp. 35-36.

c. *The wave-length of the spectrum light and the factors which influence the selectiveness of response to wave-length.* Sale¹²⁶ was the first to report that selenium is selective in its response to wave-length. He found the greatest change in resistance was caused by red light near the end of the solar spectrum; next by red of shorter wave-length; then in order by orange, green, blue and violet. Adams and Day¹²⁷ found the

¹¹⁹ Townsend, F. *Loc. cit.*

¹²⁰ Nicholson, P. J. *Op. cit.*, p. 9.

¹²¹ Marc, R. Z. *Anorg. Chem.*, 1903, 37, pp. 459-475.

¹²² Grantham, G. E. *Phys. Rev.*, 1914, 4, Ser. 2, pp. 259-266.

¹²³ Pfund, A. H. *Phil. Mag.*, 1904, 7, Ser. 6, pp. 26-39.

¹²⁴ Pfund, A. H. *Phys. Rev.*, 1912, 34, p. 370.

¹²⁵ Brown, F. C. *Phys. Rev.*, 1911, 33, pp. 14-15.

¹²⁶ Sale. *Proc. Roy. Soc. of London*, 1873, 21, pp. 283-285.

¹²⁷ Adams and Day. *Op. cit.*, p. 317.

greatest light effect in the greenish yellow, next in the red and least in the violet. Pfund¹²⁸ used lights equalized in energy by a thermopile and later by a radiomicrometer. He got the maximum response near $.7\mu$. This maximum was not changed when selenides of lead, mercury, copper, and silver were introduced. Brown and Sieg¹²⁹ found the curve of response to wave-length to vary for different types of cell. With reference to selectiveness of response to wave-length two sorts of investigation have been made, —one to determine the factors that influence this selectiveness in a given cell; the other to determine the factors which influence selectiveness in different cells.

(1) *Factors which have been found to influence the selectiveness of response for a given cell.* The following factors have been found to influence the selectiveness of response in a given cell.

(a) *Intensity.* Pfund¹³⁰ found the sensitivity curve for wave-length to vary with the intensity of the incident light. This was confirmed by Brown and Sieg¹³¹ and by Nicholson.¹³² With reference to the changes in the sensitivity curve Pfund contributes the following formula: $d=DI^\beta$ where d = change of conductivity; I = energy of illumination; and D and β are constants dependent on the wave-length of the incident light. For exposures of 12.5 sec., he found D to be constant for any particular wave-length; β was very nearly $\frac{1}{2}$ for regions of the spectrum from the violet to the yellow; but its value increased as red was approached, equalling 1 for deep red and infra-red. Nicholson verified both the formula and the constants for an exposure time of 12.5 sec.

(b) *Length of exposure time.* Nicholson,¹³³ however, found the formula contributed by Pfund to hold only for an exposure time of 12.5 sec. For longer and shorter exposures, the value of β changed. For 10 sec. exposure, it increased and the region in

¹²⁸ Pfund, A. H. Phil. Mag., 1904, 7, Ser. 6, pp. 26-39; Phys. Rev., 1909, 28, pp. 324-336.

¹²⁹ Brown, F. C. and Sieg, L. P. Phys. Rev., 1914, 4, Ser. 2, pp. 48-61.

¹³⁰ Pfund, A. H. Phys. Rev., 1909, 28, pp. 324-336.

¹³¹ Brown, F. C. and Sieg, L. P. Phys. Rev., 1913, 2, Ser. 2, pp. 487-494.

¹³² Nicholson, P. J. Phys. Rev., 1914, 3, Ser. 2, pp. 1-24.

¹³³ Nicholson, P. J. Loc. cit.

which β becomes 1 shifted towards the shorter wave-lengths. With longer exposures (15 to 20 sec.), the value of β decreased. With unlimited exposures or until a steady state of resistance of the selenium was attained, β equalled 0.5 throughout the spectrum, except at about 600μ where it equalled only 0.4. This change in the value of β for different wave-lengths with exposure time is probably in accord with Nicholson's further demonstration that selenium has a different inertia of response for different wave-lengths. This is particularly marked for the red and infra-red of the spectrum. Brown and Sieg¹³⁴ also note a change in the shape of the sensitivity curve for exposures of 30 and 0.4 sec.

(c) *Temperature, humidity and voltage.* That the selectiveness of response of selenium to wave-length varies with the temperature of the cell is mentioned by Marc¹³⁵ and Nicholson,¹³⁶ Marc finds it to vary also with the intensity of the current used; and Nicholson with the humidity.

(d) *Photo-electric currents.* Minchin¹³⁷ using seleno-aluminium cells, found that the intensity of electromotive force produced by the action of light on the cell varies with the wave-length of the incident light. It is greatest in order for yellow, orange, green, red and blue.

(2) *Factors which have been found to influence the selectiveness of response to wave-length in different cells.* The main cause of difference in the selectiveness of response to wave-length from cell to cell is according to Brown and Sieg¹³⁸ and to Dieterich¹³⁹ the temperature at which the cell was made and annealed. In general there are two groups of cells,—those with the maximum response at wave-lengths greater than 640μ ; and those with the maximum at a wave-length less than this. Cells of the former group are produced by annealing at lower temperatures, e.g., annealing at 170° C. gives a pronounced red maximum;

¹³⁴ Brown and Sieg. Phys. Rev., 1913, 2, Ser. 2, pp. 487-494.

¹³⁵ Marc, R. Z. Anorg. Chem., 1903, 37, pp. 459-475.

¹³⁶ Nicholson, P. J. *Loc. cit.*

¹³⁷ Minchin, G. M. Phil. Mag., 1891, 31, Ser. 5, pp. 207-238.

¹³⁸ Brown, F. C. and Sieg, L. P. Phys. Rev., 1914, 24, Ser. 2, pp. 48-61.

¹³⁹ Dieterich, E. O. Phys. Rev., 1914, 4, Ser. 2, pp. 467-476.

those of the latter group by annealing at high temperature, *e.g.*, at 210° C. By partial annealing at 210° and completing the process at a lower temperature, the maximum response is given in the blue and a secondary maximum in the red. Brown¹⁴⁰ confirms this result with his selenium "crystal forms" produced by the sublimation of the vapor either in a high vacuum or at atmospheric pressure. Among these forms he finds types which give the different wave-length sensitivity curves found by Dieterich in the different cells annealed at the various temperatures. Brown believes that there are at least three forms of metallic selenium of widely different electrical resistivity. These forms are produced at different temperatures. That is, at high temperatures, for example, crystals of maximum sensitivity to red light are not allowed to form.

d. *The intensity of white light.* Attempts have been made to use the selenium cell both as a radiometer and a photometer. In the latter case, the following laws of change of resistance with change of intensity have at different times been formulated. When m = conductivity, i = light intensity, R = resistance, and the other quantities are constants, Rosse,¹⁴¹ Adams and Day,¹⁴² and Berndt¹⁴³ give the formula $i = cm^2$; Hopius,¹⁴⁴ $i = cm^3$; Athanasiadis,¹⁴⁵ $i = m(m-a)b$; Hesehus,¹⁴⁶ $i = b^m - 1$; Ruhmer,¹⁴⁷ $R_0/R_b = (b/a)^a$; Stebbins, $i = cm$. (See Brown, *Phys. Rev.*, 1911, 33, pp. 1-26.) Brown states that although the illumination used, time of exposure, and construction of cell varied in the work of the above men, it appears obviously futile from these results to look for a universal law of conductivity for the selenium cell as a function of intensity of illumination.

In summarizing the difficulties that apply to the use of the

¹⁴⁰ Brown, F. C. *Phys. Rev.*, 1914, 4, Ser. 2, pp. 85-98; see also Dietrich, *ibid.*, p. 474.

¹⁴¹ Rosse. *Phil. Mag.*, 1874, 47, Ser. 4, pp. 161-164.

¹⁴² Adams and Day. *Op. cit.*, p. 318.

¹⁴³ Berndt, G. *Phys. Z.*, 1904, 5, pp. 121-124.

¹⁴⁴ Hopius. *Jurn. Russk. Fizik. Chimicesk. Obscestva*, 1903, 35, pp. 581-585.

¹⁴⁵ Athanasiadis, G. *Ann. der Physik*, 1908, 25, pp. 92-98.

¹⁴⁶ Hesehus, N. A. *Jurn. Russk. Fizik. Chimicesk. Obscestva*, 1905, 37, pp. 221-231; *Phys. Zeit.*, 1906, 7, pp. 163-168.

¹⁴⁷ Ruhmer, E. *Phys. Zeit.*, 1902, 3, pp. 468-474.

selenium cell in the measurement of light, the following points then may be noted. (1) The fatigue effects and the effects of previous exposure to light are so great that it is exceedingly difficult to keep the cell in a state of constant sensitivity (2) The amount of response is not only a function of the time of exposure to the light, but apparently rather complexly so. (3) There is not only selectiveness of response to wave-length but the amount of this selectiveness varies with the intensity of light, with the strength of the battery current, with the temperature of the cell and with humidity.¹⁴⁸ While there is a possibility of controlling the last three of this latter group of factors, there seems no way to deal satisfactorily for any wide use of the cell with the first, or what may be termed roughly a "Purkinje phenomenon." Because of this factor a calibration of the cell for wave-length for one intensity of light would not hold for all intensities, which would limit the use of the cell to the intensity of light for which it was calibrated or for ranges for which there is no change in relative sensitivity to wave-length. That is, any wide use of the cell would require both a wave-length and an intensity calibration in terms, for example, of the responses of the non-selective instruments. And (4) there seems to be no regular relation of the amount of response to the amount or intensity of light used even when the lights are of the same composition. At least according to Brown this is the conclusion that must be drawn from the work that has been done with white light. If this be true, the possibilities of use of the selenium cell as a radiometric instrument seem in general practice to be limited to the equalization of light intensities and this, unless correction factors are used, only in case the lights are of the same composition. In this regard its case is similar to that of the eye when considered as a possible radiometric instrument.¹⁴⁹

¹⁴⁸ It will be remembered also that the intensity of the photo-electric currents that are set up by the action of light on selenium which are an important factor in the variability of action of the cell, are different for the different wave-lengths. This factor obviously can not be controlled and there seems no satisfactory calibration for it.

¹⁴⁹ In this lack of a simple relation between the amount of intensity of light and amount of response, the selenium cell again presents an analogy to

4. *Theories of the action of light on selenium.* It may be of interest to append here a brief account of the theories that have been advanced to explain the action of light on selenium. The change of resistance of selenium under the action of light has at various times been thought to be a heating effect, to be electrolytic in nature, to be electronic, or to be of chemical origin. The first view was held by Sale, Sabine and Moser. Sabine, for example, thought that the action is similar in character to that of a dielectric "more or less charged with conducting crystals." In such a case the light by its heating effect would modify the surface tension of the selenium, which modification would probably cause an expansion of its crystalline surface and this in turn would result in a closer contact among the superficial crystals. This view was disproved by the demonstration that the light effect of the different wave-lengths on selenium does not correlate with their heating effect.

The theory that the action is electrolytic was first proposed by Adams and Day. They did not claim that actual electrolysis takes place, however, but that the molecular structure or crystalline condition of the selenium is altered or modified by the action of a current of electricity in such a manner as to produce effects analogous to those which would occur if the selenium were an electrolyte and were actually decomposed by the current. Furthermore, they thought that the action of light falling on selenium is to promote crystallization and thus to diminish its resistance to an electric current, inasmuch as in changing to the crystalline state selenium becomes a better conductor of electricity. And as this crystallization is greatest in the exterior layers of the

the eye. And as in case of the eye this relation has as yet proven incapable of mathematical formulation. Fechner in his attempt to give a mathematical expression of the relation between stimulus and response for sensation in general was only trying to do what a number have tried to do with regard to this reaction both of the selenium cell and the photographic plate. When one knows how signally the attempt to find an expression separately for either of these reactions has failed, one realizes still more clearly the *a priori* improbability of finding a single expression that will apply to the reactions of five sensory mechanisms so differently constituted as they seem to be.

selenium, a flow of energy from within outwards is produced which under certain circumstances appears to produce an electric current (the photo-electric current). Bidwell thought the action is really electrolytic, impure selenides having been used or selenides having been formed between the selenium and the metallic electrodes.

Bidwell's view was disproved by Pfund and later by Berndt, both of whom used purified selenium and carbon electrodes and got greater sensitivity to light with the purified than with the impure selenium. Pfund developed an electronic theory of the action, an explanation that had previously been suggested by Nagaoka. He considered the effect due to a resonance of the electrons in the atom under the action of light, causing explosions which lead to an increase in the number of conducting electrons. There is, moreover, a "critical depth" of penetration above and below which the action on selenium is less pronounced. This fact accounts for the selective response to wave-length of light, the maximum response being to that light which penetrates to the "critical depth"; also for the change in this selectiveness of response to wave-length with change of intensity of the incident light. This view is held also by Ries and Nicholson.

The chemical theory has been followed among others by Marc, Montén, Kruyt, Pochettino and Berndt. These men have obtained evidence that leads them to believe that there exist at least two forms of metallic selenium of widely different electrical resistivity, and they assume that illumination brings about a transformation from the less to the more conductive of the two.

Brown claims to have found and isolated three forms of selenium crystals. They are produced at different temperatures in the annealing process and possess different conductivity. These "crystal forms" have also a different selectivity of response to wave-length. In his opinion the character of the conductivity curves for the four known varieties of light-sensitive selenium can be explained by assuming the existence of three components in dynamic equilibrium under a given illumination, temperature, pressure and electrical potential difference. Any agency that changes

the conductivity of selenium is of such a nature that it alters the rate of interchange between these components.

F. THE PHOTO-ELECTRIC CELL.

The action of the photo-electric cell depends upon the effect of light on the capacity of certain metals to hold a negative charge of electricity. Knowledge of the action of light on the conduction of electricity goes back to the discovery by Hertz in 1887 that the incidence of ultra-violet radiations on a spark gap facilitates the sparking. This led to a general investigation of the effect of light on the conduction of electricity.¹⁵⁰ The discoveries which paved the way directly for the invention of the photo-electric cell were those pertaining to the effect of light on the electrical condition of certain metals. It was found, for example, that a zinc plate exposed to light becomes slightly positively charged; that a negatively charged plate becomes less negatively charged; and that a positively charged plate is not affected. Later studies showed that the electrical condition of all metals is changed to some extent by the action of light. Those affected most are, according to Elster and Geitel,¹⁵¹ rubidium, potassium, alloy of potassium and sodium, sodium, lithium, magnesium, thallium, and zinc.

The essential parts of the photo-electric cell are as follows. There must be juxtaposed in a glass tube or vessel a negatively charged surface of the metal in question (the cathode) and a conductor or anode to receive the charge as it is lost by the cathode under the action of light. Connected in series with the cell is either an electrometer or a galvanometer. In some cases the inside of the cell is coated with the metal forming the cathode and a receiving wire is suspended in the cell. In others, the negatively charged metal is suspended in the cell and the body of the tube silvered on the inside serves as anode. The photo-

¹⁵⁰ See Hallwachs, W. *Ann. der Phys.*, 1888, 33, p. 301; Hoor, M. *Reper-tonnière des Physik*, 1889, 25, p. 91; Righi, A. *Comptes Rendus*, 1888, 106, p. 1349 and 107, p. 559; Stoletow, A. *ibid.*, 1888, 106, p. 1149, 1593 and 107, p. 91; 1889, 108, p. 1241; *Physikalische Revue*, Stuttgart, 1892, 1.

¹⁵¹ Elster, J. and Geitel, H. *Nature*, 1894, 50, p. 451.

electric current may be measured in five ways: (1) by the rate of drift of an electrometer needle; (2) by the ballistic method or the measurement of the charge acquired in a definite exposure time by an electrometer connected with the cell; (3) by measuring the potential across the terminals of a high resistance in series with the cell; (4) by balancing the photo-electric current with a current variable in a known manner, by means of either an electrometer or a sensitive galvanometer; and (5) by the deflections of a sensitive galvanometer. Ives¹⁵² commenting on these recommends the third method. He finds the first inadvisable because the rate of drift is not uniform; the second, because the deflection varies with the exposure time; and the fifth, because it is insensitive.

1. *Factors that have been taken into account in the construction and use of photo-electric cells.* The following are some of the factors that have been taken into account in the construction and use of sensitive cells.

a. *The metal used.* The metal used should have a high emissive power and should permit of a certain ease in handling. Different metals have been used by different experimenters. Compton and Richardson,¹⁵³ for example, used aluminium, platinum, sodium and caesium. Potassium and sodium have been most frequently employed. For a summary of the metals used by different investigators, see Allen, Photo-electricity, 1913, p. 68.

b. *The residual gas.* It is desirable of course to have for the residual gas one which ionizes easily but not to such an extent that the recombination of the ions is measurable, also one which is easy to handle. Elster and Geitel¹⁵⁴ tested the rate of loss of charge from an illuminated surface through air, carbon dioxide, oxygen and hydrogen; and found that the rate of leak through carbon dioxide was much faster than for any of the other gases. Hydrogen, helium or argon have been most frequently used in the more recent work with photo-electric cells.

¹⁵² Ives, H. E. *Astrophys. Jour.*, 1914, 39, p. 428.

¹⁵³ Compton, K. T. and Richardson, O. W. *Philos. Mag.*, 1913, 26, Ser. 6, p. 561.

¹⁵⁴ Elster, J. and Geitel, H. *Ann. der Phys.*, 1890, 41, p. 166.

c. *The pressure of the residual gas.* There are two factors here which work against each other. If we consider the current from the cathode to the suspended loop of wire as the discharge of negative electrons from the cathode, a vacuum would offer the least impedance. An advantage is gained, however, by adding to the electrons sent off by the metal, electrons freed by ionizing the gas in the tube. The effect of pressure on the intensity of the photo-electric discharge has been investigated by Stoletow,¹⁵⁵ Schweidler¹⁵⁶ and Kemp¹⁵⁷ by comparing the intensity of the current for different pressures under otherwise constant conditions. For Stoletow the most favorable range of pressures is from 0.275-2.48 mm.; for Schweidler, from 1-2 mm.; for Kemp, from 2-3 mm. In a recent work Ives, Dushman and Karrer¹⁵⁸ found that the pressure giving the greatest sensitivity varies with the voltage; further that the "photo-electric sensitiveness does not disappear when the metal is made as gas-free as possible and the degree of vacuum is made as high as possible."

d. *The potential difference between anode and cathode.* There must be a sufficient difference of potential between anode and cathode to guarantee that all of the electrons are drawn to the anode, *i.e.*, there must be a saturation difference of potential. The difference must not, however, be so great as to cause sparking, and it must be kept fairly constant. From about 20 to 180 volts are generally used. For an investigation of this effect, see Stoletow¹⁵⁹ and Schweidler.¹⁶⁰

e. *The galvanometer or electrometer.* The electrometer is more sensitive than the galvanometer, but it is so much harder to manage that it seems to be the general opinion that it should be used only for measuring very low intensities.¹⁶¹

¹⁵⁵ Stoletow, A. *Comptes Rendus*, 1888, 107, p. 91; *J. de Phys.*, 1890, 9, p. 468.

¹⁵⁶ Schweidler, E. R. *Sitzungsber. der Wien. Akad.*, 1898, 107, 2a, p. 881.

¹⁵⁷ Kemp, J. G. *Phys. Rev.*, 1913, 1, Ser. 2, p. 274.

¹⁵⁸ Ives, Dushman and Karrer. *Astrophys. Jour.*, 1916, 43, p. 9.

¹⁵⁹ Stoletow, A. *Comptes Rendus*, 1889, 108, p. 1241.

¹⁶⁰ Schweidler, E. R. *Loc. cit.*

¹⁶¹ See Richtmyer, F. K. *Transactions Illuminating Engineering Society*, 1913, 8, p. 461.

f. *The angle of the incident light.* Elster and Geitel¹⁶² first reported that the angle at which the light strikes the cathode plate causes a difference in the effect on the plate. This effect has been investigated among others by Pohl¹⁶³ and Kunz.¹⁶⁴ Elster and Geitel claim to have overcome the effect due to angle of incidence by using a diffusing screen so that the light falls equally at all angles on the plate.¹⁶⁵ Pohl and Pringsheim¹⁶⁶ have found in addition that the curve of selective response to wave-length of the photo-electric cell varies with the angle of incidence.

g. *Dark effects and after-effects.* When a cell is charged and left in the dark, a slight leakage or discharge is found to take place which is increased following an exposure to light. This leakage Elster and Geitel¹⁶⁷ believe to be due to a conduction of current over the surface of the glass from the cathode to the anode circuit. In any event guard rings of metal connected to earth, placed on the inside and outside of the tube between the cathode and anode circuit, completely eliminated the leakage.

h. *Fatigue effects.* Some metals when freshly prepared throw off many electrons when acted upon by light, then the number becomes less. This decrease in the responsiveness of the metal is apparently rapid at first, then becomes slower, ceasing perhaps in a few days. Allen,¹⁶⁸ for example, measured the photo-electric activity at different intervals from 2 to 100 minutes after polishing the surface of a zinc plate. He found a decrease in activity which was rapid for the first few minutes, then more gradual after 20 to 30 minutes. Sadzewicz¹⁶⁹ reports a similar result. The effect has also been investigated by Holman,¹⁷⁰ Hallwachs,¹⁷¹

¹⁶² Elster, J. and Geitel, H. Wied. Ann. der Phys., 1894, 52, p. 433; 1895, 55, p. 684; 1897, 61, p. 445.

¹⁶³ Pohl, R. Phys. Z., 1909, 10, p. 542.

¹⁶⁴ Kunz, J. Phys. Rev., 1909, 29, p. 174.

¹⁶⁵ Elster, J. and Geitel, H. Phys. Z., 1912, 13, p. 740.

¹⁶⁶ Pohl, R. and Pringsheim, P. Deutsch. Phys. Gesell., 1910, 12, p. 215.

¹⁶⁷ Elster, J. and Geitel, H. Phys. Z., 1913, 14, p. 741.

¹⁶⁸ Allen, H. S. Proc. Roy. Soc., 1907, 78, Ser. A., p. 483.

¹⁶⁹ Sadzewicz, M. Acad. Sci. Cracovie, Bull., 1907, 5, p. 497.

¹⁷⁰ Holman, W. F. Phys. Rev., 1907, 25, p. 81.

¹⁷¹ Hallwachs, W. Ann. der Phys., 1907, 23, (4), p. 459.

Compton and Richardson,¹⁷² Buisson,¹⁷³ Ladenburg¹⁷⁴ and Bergowitz.¹⁷⁵ Bergowitz claims, however, that there is no fatigue in case of cells whose negative poles are formed of alkali metals. This statement is confirmed by Elster and Geitel¹⁷⁶ who say: "sogenannte 'Ermüdungserscheinungen' an Alkalimetallzellen nicht auftreten." Ives, Dushman and Karrer, however, apparently found fatigue effects in potassium cells.

2. *Comparative advantages and disadvantages of the photo-electric cell.* The photo-electric cell has the advantage of comparatively high sensitivity. To offset its sensitivity, however, it has the following serious disadvantages. (a) It is selective in its response to wave-length. The shorter wave-lengths are overestimated. This selectiveness moreover, varies with the metal used in the cell. Pohl, and Pohl and Pringsheim have plotted the curves of response to wave-length for the following metals: mercury,¹⁷⁷ platinum and copper,¹⁷⁸ potassium-sodium alloy,¹⁷⁹ rubidium, potassium and sodium,¹⁸⁰ barium, ¹⁸¹ lithium and sodium,¹⁸² magnesium and aluminium,¹⁸³ and calcium.¹⁸⁴ The selectiveness for calcium is found to be very similar to the selectiveness of the eye to wave-length. Richtmyer has plotted the curve for sodium;¹⁸⁵ Hallwachs for potassium;¹⁸⁶ Kunz for sodium-potassium alloy;¹⁸⁷ and Elster and Geitel for

¹⁷² Compton, K. and Richardson, O. *Philos. Mag.*, 1913, 26, Ser. 6, p. 561.

¹⁷³ Buisson, A. *Comptes Rendus*, 1900, 130, p. 1298; *Ann. Chim. Phys.*, 1901, 24, p. 320.

¹⁷⁴ Ladenburg, E. *Ann. der Physik*, 1903, 12, p. 558.

¹⁷⁵ Bergowitz, K. *Phys. Z.*, 1907, 8, p. 373.

¹⁷⁶ Elster and Geitel, H. *Phys. Z.*, 1913, 14, footnote p. 742.

¹⁷⁷ Pohl, R. *Phys. Gesell. Verh.* 1909, 11, p. 609.

¹⁷⁸ Pohl, R. *Ibid.*, 1909, 11, p. 339.

¹⁷⁹ Pohl, R. *Ibid.*, 1909, 11, p. 715; Pohl and Pringsheim; *Ibid.*, 1910, 12, p. 215, 349, 682, 697.

¹⁸⁰ Pohl and Pringsheim. *Ibid.*, 1910, 12, p. 1039; 1911, 13, p. 219.

¹⁸¹ Pohl and Pringsheim. *Ibid.*, 1911, 13, p. 474.

¹⁸² Pohl and Pringsheim. *Ibid.*, 1912, 14, p. 46.

¹⁸³ Pohl and Pringsheim. *Ibid.*, 1912, 14, p. 546.

¹⁸⁴ Pohl and Pringsheim. *Ibid.*, 1913, 15, p. 111.

¹⁸⁵ Richtmyer, F. K. *Phys. Rev.*, 1910, 30, p. 385.

¹⁸⁶ Hallwachs, W. *Ann. der Phys.*, 1909, 30, (4), p. 593.

¹⁸⁷ Kunz, J. *Phys. Rev.*, 1909, 29, p. 212.

rubidium, potassium and sodium.¹⁸⁸ (b) Griffith¹⁸⁹ working with ultra-violet light, and Dember¹⁹⁰ working with the visible spectrum, both claim that it is also selective in its response to intensity. That is, they do not find a constant relation between intensity of radiation and photo-electric current. Elster and Geitel,¹⁹¹ however, found on the other hand a constant relation between intensity of light and the response of the cell except in case of very intense light. As source they used the light of the sun, a mercury arc, a Nernst lamp of 32 cp. and a 2-volt carbon lamp. A variable resistance was used with the 2-volt lamp; also in a part of the work its light was passed through a blue filter. Richtmyer¹⁹² found the photo-electric current from a sodium surface under the action of light from an incandescent lamp to be proportional to the intensity of the incident light for very low intensities (0.007 candle-foot) up to 620 candle-feet. Ives in 1914 found that the relation between illumination and photo-electric effect is not linear and differs from cell to cell. In 1916, however, working with Dushman and Karrer he reports that the cause of these varied relationships lies in "focusing effects." "By this term is meant a change of direction of the electron stream as the number emitted changes, whereby a different proportion of the whole number of electrons reaches the receiving electrode" (p. 25). For the elimination of these effects he recommends a cell of special construction having absolutely no free surfaces on which electric charges can collect (see p. 30). This cell, he finds, gives a rectilinear relationship between intensity of illumination and photo-electric effect.¹⁹³ (c) The cell is not sensitive to

¹⁸⁸ Elster and Geitel. *Ann. der Phys.*, 1894, 52, (3), p. 438.

¹⁸⁹ Griffith, I. *Phil. Mag.*, 1907, 14, (6), p. 297.

¹⁹⁰ Dember, H. *Ber. d. kgl. sächs. Akad. d. Wiss.*, 1912, 64, p. 266.

¹⁹¹ Elster and Geitel. *Phys. Z.*, 1913, 14, p. 741; 1914, 15, p. 610.

¹⁹² Richtmyer, F. K. *Phys. Rev.*, 1909, 29, p. 71, 404.

¹⁹³ Kunz recently reports an investigation of cells of special construction the responses of which to white light for a wide range of intensities deviates so little from a linear function of intensity of light as to permit of the use of the cell for many photometric purposes. He verifies Ives's claim that in case of the older spherical form of cell, the photo-electric current is not proportional to the intensity of light. He found also that the Talbot-Plateau law holds for the responses of the cells described. (*Astrophysical Jour.*, 1917, 45, pp. 69-88.)

heat radiations, hence can not be calibrated against the total of radiation of a black body. The value of its responses in terms of energy units can be determined feasibly and conveniently only by the aid of some other radiometric instrument which responds to the total of radiation, *e.g.*, the thermopile, the bolometer, the Nichol's radiometer, etc. (d) With the present knowledge and control of the factors which influence its sensitivity, the cell can scarcely be recommended as giving results with a degree of reproducibility which is entirely satisfactory. Thus from the standpoint both of reproducibility and selectiveness of response the use of the cell in its present stage of development even as an energy comparator of the visible radiations can scarcely be considered as advised by the radiometric specialist. However, the cell is still being developed and perfected and may yet be of service in measuring light intensities.

As a light measuring instrument the photo-electric cell has the following advantages. (a) It has a comparatively high sensitivity especially to the shorter wave-lengths.¹⁰⁴ And (b) it responds very quickly to the light stimulus. Richtmyer claims, for example, that one of the special fields in which the cell promises to be serviceable is for exposures too short for the eye to be used with accuracy and convenience.¹⁰⁵

G. THE PHOTOGRAPHIC PLATE.

1. *The blackening of the plate and the factors which influence this action.* When light acts on a photographic plate a chemical change takes place in the sensitive film which renders it opaque to light when the plate has been developed. This is called the blackening of the plate, and is the response that must be calibrated if the plate is to be used as a light-measuring instrument. Unlike

¹⁰⁴ The shape of the curve representing the photo-electric response to wave-length is for calcium very similar to that of the eye. The maximum for most of the other metals that have been investigated, occurs much nearer to the violet end of the spectrum. The position of this maximum, as has been stated above, depends on the kind of metal used to give the photo-electric effect.

¹⁰⁵ Richtmyer claims advantage for the cell only in certain special fields of work. (See. Trans. Illuminating Eng. Soc., 1913, 8, p. 459.)

the thermopile, for example, which gives its maximum response when once a thermal equilibrium has been attained and is from then on practically independent of the time of exposure to the light, the blackening of the photographic plate is a function of the time of exposure to the light as well as of its intensity and wavelength. An important problem in calibrating has been, therefore, to find out whether the amount of blackening sustains any regular relation to these two factors. If so, the relation can be expressed in terms of a formula; if not, the calibration must at every point be empirical. With regard to the degree of regularity of the blackening there has been a great deal of disagreement. In 1862, for example, Bunsen and Roscoe announced that the blackening may be expressed by the formula $S = i t$, in which S represents the blackening, i the intensity of light and t the time of its action on the plate.¹⁹⁶

¹⁹⁶ That equal amounts of blackening are always produced by equal intensities of light and equal times of exposure was first accepted as a theoretical principle by Malaguti (*Ann. de Chemie et de Phys.*, 26, p. 5; *Pogg. Ann. der Phys.*, 1840, 49, p. 567).

It was first experimentally demonstrated within a narrow range of light intensities (1 to 2½ in an arbitrary scale) by Hankel (*Abhandl. der k. Sächs. Gesell. d. Wiss. z. Leipzig*, 1864, 9, p. 55). Its first claim to establishment as a general law came from Bunsen and Roscoe (*Annal. der Phys. und Chemie*, 1862, 117, pp. 529-562). They say: "So wird man den Satz als feststehend betrachten dürfen, dass innerhalb sehr weiter Gränzen gleichen Producten aus Lichtintensität und Insulationsdauer gleiche Schwärzungen auf Chlorsilberpapier von gleicher Sensibilität entsprechen." The law may be applied under the following conditions. "(1) Wenn die bei Messungen des gesammten Himmelslichts in Betracht kommenden Lichtstärken nur noch von so kurzen Inductionsphänomenen begleitet sind, dass die dadurch erzeugten Störungen innerhalb der erlaubten unvermeidlichen Beobachtungsfehler fallen; (2) wenn es möglich ist, eine photographisch sensible Schicht von völlig constant Empfindlichkeit darzustellen; (3) wenn sich eine unverändliche, zu jeder Zeit und an jedem Orte leicht wieder hervorzubringende Schwärzung finden lässt, die eine sichere Vergleichung mit einer photographisch geschwärzten Fläche zulässt."

The first suggestion that photographic action may be used as a means of measuring light intensities seems to have come from Sir J. F. W. Herschel who describes an "actinograph or self-registering photometer for meteorological purposes" in Section VIII of paper "On the Chemical Action of the Rays of the Solar Spectrum on Preparations of Silver and other Substances, both metallic and non-metallic, and on some Photographic Processes (*Philos. Trans.*, 1840, 130, pp. 1-61); and suggests on pp. 46-47 that the photographic

This statement that the effect of the light on the plate may be expressed by the product of the intensity and time of exposure or is directly proportional to the energy of the light incident on the plate, has become widely known in the subjects of Chemistry and Physiology as the Bunsen-Roscoe law. If it were true that the blackening is proportional to the energy falling upon the sensitive surface, the photographic plate could be used directly as an energy measurer and would serve as an exceedingly useful means of measuring light intensities, for it has the additional advantages of great sensitivity and of integration of action through an interval of time. More recent investigations of the blackening, however, beginning with Abney in 1874,¹⁹⁷ give little support to the law formulated by Bunsen and Roscoe. The blackening does not vary regularly with the intensity and time of exposure for any considerable range of intensities and times of exposure. It is also, as is well-known, selective in its response to wave-length and this selectiveness varies with the intensity of the light. That is, like the eye the photographic plate is selective in its response to intensity (a crude analogy to the Purkinje phenomenon) and is irregular in its action through an interval of time.¹⁹⁸ Moreover,

plate may be used to measure light. Later A. Claudet describes the "photographometer, an instrument for measuring the intensity of the chemical action of the rays of light on all the photographic preparations, and for comparing with each other the sensitiveness of these different preparations" (*Philos. Mag.*, 1848, Ser. 3, 33, pp. 329-335). Claudet also refers to T. B. Jordan who invented an instrument which he called a heliograph consisting of a cylinder covered with sensitized paper placed parallel to the axis of the ecliptic, which turned to follow the sun. The object of this apparatus was to get the actinic value of the sun's rays at different times in the day. The instrument was improved by R. Hunt in 1845 and called by him an actinograph. In the work of these men we find the somewhat obscure beginnings of the subject of actinometry which is later to compete with photometry and radiometry as methods of measuring light intensities.

¹⁹⁷ Abney, W. deW. *Philos. Mag.*, 1874, 48, Ser. 4, pp. 161-165.

¹⁹⁸ In addition to the above three characteristics crudely analogous to the eye, the action on the photographic plate is said by Bunsen and Roscoe to show an inertia or lag in coming to its full value. (See Pogg. *Annal.*, 1855, 100, p. 481-516; also *Photochem. Untersuch.*, Ostwald's *Klassiker*, 34, p. 363.)

The analogy in the response of the photographic plate to the optical Purkinje phenomenon was first mentioned by Miethe, *Zur Aktinometrie astronomisch-photographischer Fixsternaufnahmen*, Göttingen, 1889. The

the response varies with other factors which are in practice difficult to control and exceedingly troublesome if not impossible to take into account in the derivation of formulae.¹⁹⁹

The tendency in fact among recent investigators has been to question whether a mathematical expression can be given to the action for any considerable range of intensities and times of exposure and to disagree widely with regard to the formula that should be used for the range of intensities and times of exposure regarded as most favorable to regularity of action. At different times the following formulae have been derived to express the action: $S = i t^p$ in which p represents a constant with a value of 0.86 (Schwarzchild);²⁰⁰ $S = k \cdot i t^p$ in which k and p are both constants (Leimbach);²⁰¹ $S = k \cdot \log. i t^p$ (Parkhurst);²⁰² and $S = \log. (k \cdot i^m t^n)$ in which k , m and n are all constants (Stark).²⁰³

The validity of the above formulae has been the subject of considerable experimental investigation. Renwick,²⁰⁴ for example,

phenomenon is discussed and tested experimentally within certain limits of difference in wave-length by Schwarzchild, (Sitzungsber. d. Wien. Akad., Math.-Naturwiss. Classe, 1900, 109, 2a, pp. 1127-1135). Discussing the influence of the optical Purkinje phenomenon on the systematic differences in the different brightness catalogues of stars, he says: "Aber es besteht auch für die Photographie ein dem Purkinje-Phänomen ganz analoger Übelstand. Zwei verschiedenfarbige Lichtquellen, die bei einer gewissen, für beide gleichen Expositionszeit gleiche Schwärzung ergeben, erfüllen diese Bedingung nicht mehr, wenn man die Expositionszeit ändert oder die Intensitäten der Lichtquellen im selben Verhältnisse verstärkt" (p. 1128).

¹⁹⁹ Blackening is said to vary with wave-length of light, with kind of plate, with process of developing, with intensity of light, with time of exposure, with temperature of plate, and with type of exposure—continuous or intermittent. With regard to type of exposure Abney (Journ. of the Phot. Soc., 1893-4, p. 63), Eder (Sitzber. d. Wiener Akad., Math.-Nat. Classe, 1899, 108, 2a, p. 1433), Englisch (Archiv. f. Wiss. Photog., 1899, 1, p. 117; 1900, 2, p. 131), and Schwarzchild (Astrophys. Journ., 1900, 11, pp. 92-100) all find a less effect with an intermittent than with a continuous exposure. Abney, for example, finds the retardation to be more pronounced the greater is the ratio of closed to open sector, the greater the speed of rotation employed, and the less the light intensity.

²⁰⁰ Schwarzchild, K. *Astrophys. Journ.*, 1900, 11, pp. 89-92.

²⁰¹ Leimbach, G. *Zeit. f. wiss. Photog.*, 1909, 7, p. 174.

²⁰² Parkhurst, J. A. *Astrophys. Journ.*, 1909, 30, p. 33.

²⁰³ Stark, J. *Annal. der Phys.*, 1911, 35, (4) pp. 461-486.

²⁰⁴ Renwick, F. F. *Photog. Journ.*

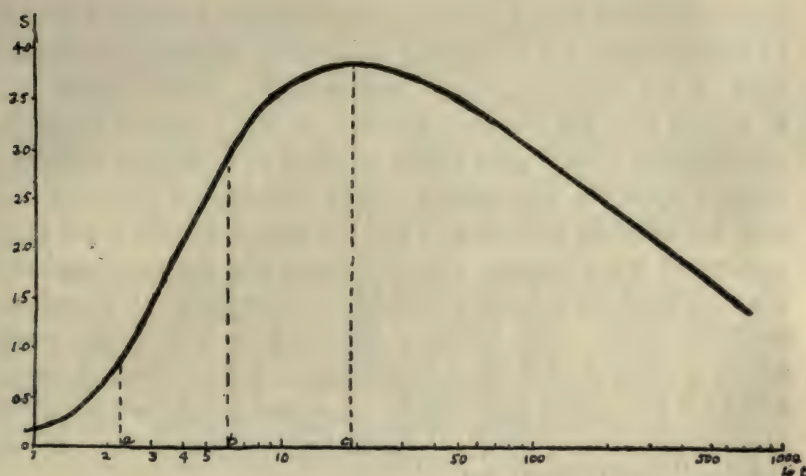


Fig. 1.

contends that the Bunsen-Roscoe law falls short about 1.17 per cent. Schwarzschild²⁰⁵ finds for p a value of 0.86; while Tikhoff²⁰⁶ claims that for the photographic rays p varies from 0.67 to 0.79; and for the green-yellow rays, from 0.91 to 0.96. Parkhurst²⁰⁷ states that p is not a constant but a variable depending for its value (a) upon the density of the image, (b) the kind of plate used, and (c) the light filter employed. Geiger²⁰⁸ finds that the law formulated by Schwarzschild is approximately correct within certain limits of time of action on the plate. Keeping the intensity the same and plotting the log. of the time against the blackening, Geiger obtains the curve given in Fig. 1. So plotted the curve should be a straight line according to the formula $S = i t^p$ if p is a constant. Between the points a and b the line is almost straight, he finds. Between these limits alone then the action is capable of approximate formulation, and the plate should not be used for light measurements for lengths of ex-

²⁰⁵ Schwarzschild, K. *Loc cit.*

²⁰⁶ Tikhoff, G. Mittheilungen d. Nikolai-Hauptsternwarte zu Pulkow, 1909, 31, (3), p. 31; Comptes Rendus, 1909, 148, p. 268.

²⁰⁷ Parkhurst, J. A. Astrophys. Journ., 1909, 30, p. 34.

²⁰⁸ Geiger, L. Annal. der. Phys., 1911, 37, (4) pp. 68-78.

posure time that do not fall within these limits. Abney²⁰⁹ conducts an investigation for the special purpose of showing the dependence of the blackening upon wave-length. Stark²¹⁰ finds that k , m and n in the formula $S = \log. (k \cdot i^m t^n)$ depend upon the wave-length; Eder²¹¹ gets results showing that the time exponent depends upon the wave-length; while Leimbach²¹² contends that both the intensity exponent m and the time exponent n are independent of wave-length. Stark²¹³ claims that the time exponent n is a constant for a range of light intensities from 1 to 1600 in an arbitrary scale. The intensity exponent m within this range varies 5 percent. for some emulsions and for others it varies widely. k also varies quite widely. k , m and n may be considered as constants over a range of intensities varying from 1 to 100 for the "normalbelichtung."

2. *The possibilities of using the photographic plate in quantitative work.* While the above may be considered only as the briefest mention of the quantitative work that has been done on the blackening of the photographic plate, still it is enough to show that the plate can scarcely be considered as a feasible light-measuring instrument. Its quickness of response, its sensitivity, and especially its integration of action through an interval of time make it very valuable, however, for many kinds of scientific work in which quantitative comparisons are not important.

H. THE EYE.

1. *The two possibilities of using the eye in light measurements.* As an instrument for the measuring or comparing of light intensities, the eye may be regarded in two ways. (1) It may be used to rate and compare lights designed for its own service. In the production of illumination effects this is the work of photometry and should be done by the eye or some instrument calibrated to give results in terms of the response of

²⁰⁹ Abney, W. deW. Proc. Roy. Sec., 1901, 68, pp. 300-321.

²¹⁰ Stark, J. Annal. der Phys., 1911, 35, (4), pp. 461-486.

²¹¹ Eder, J. M. *Op. cit.*, p. 1473.

²¹² Leimbach, G. Diss. Göttingen, 1909; Zeit. f. wiss. Photog., 1909, 7, p. 257.

²¹³ Stark, J. *Loc. cit.*

the eye. And (2) it may be used in balancing the energies of lights of the same spectro-radiometric composition. Used as such it is one of the most sensitive of the energy-comparing instruments. It can not be used, however, to balance radiometrically lights differing in composition without elaborate calibration, because of the degree of selectiveness of its response to wave-length.²¹⁴

2. *The comparative advantages and disadvantages of the use of the eye for balancing energies of light of the same spectro-radiometric composition.* When used to balance lights of the same spectro-radiometric composition, the eye has the following advantages. (a) It is highly sensitive, among the most sensitive of the light measuring instruments. (b) It is quick in its action, reaching its maximum of response in times variously esti-

²¹⁴ It is in fact our interest in making a quantitative determination of the selectiveness of this response both to wave-length and to intensity in all the ways in which the eye responds to its stimulus, that has led us to attempt to help bring about means of rendering energy measurements feasible for the work in psychological optics. Under this heading would come, for example, the investigation of the selectiveness of the eye's achromatic response both to wave-length and to intensity with its wide application to photometry and light specification; the selectiveness of the chromatic response; the selectiveness to wave-length and intensity shown in the rise and decay of both types of response; the selectiveness found in after-image and contrast response; etc. In fact neither the characteristics and possibilities of the eye as a measuring instrument nor its peculiarities as a sense organ can be definitely known without a common unit in terms of which to evaluate the different wave-lengths to which it gives response. It is, moreover, obvious that neither the unit nor method of measurement must in any way involve the peculiarities of the responses of the eye itself.

In this work our point of view is to investigate the responses of the eye just as the physicist investigates the responses of his instruments. Too frequently this investigation has received its direction from theories and doctrinal conceptions. Such investigations can not help but be narrow in their scope and are moreover apt to lead to wrong conclusions. Much more will be accomplished, we believe, by holding theoretical and doctrinal interests in abeyance for a time and to approach the study of the eye's responses with the broader purpose of finding out what they are from the purely descriptive point of view, using methods and technique designed for such a purpose and not for the confirmation or destruction of a theory. In any event the two points of view should be kept separate in the work of investigation, and in the evaluation of results it should be clearly recognized to what degree a result is the product of the method of working employed.

mated from 0.014 to 0.541 sec. And (c) it possesses great advantages in the ease and convenience with which it may be used. Its disadvantages are very similar to those of the photo-electric and selenium cells. (a) Like both of these instruments it is very selective in its response to wave-length and can not be used, therefore, as an energy comparator of lights differing in spectro-radiometric composition without elaborate calibration. (b) Like the selenium cell and photographic plate it is selective also in its response to intensity. It is perhaps more selective in this regard than the selenium cell. And (c) it responds only to the visible spectrum and can, therefore, be calibrated against a black body only with exceeding difficulty and with many chances of error both in the calibration and its subsequent use. (See preface, pp. vi-vii. If calibrated, the thermopile or some other instrument which is sensitive to the total of radiation would ordinarily be employed and the calibration be made in terms of its responses. The work of calibration presents, moreover, great difficulties in case of the eye because of the lack of a fixed or closely reproducible scale of responses capable of numerical rating which can be correlated with the responses of the calibrating instrument. Two possibilities for calibrating suggest themselves: (a) the determination of sensitivity curves, valid only for the particular eye, the exact physical and physiological conditions, etc. for which the calibration was made, and for ranges of intensity in which no changes in relative sensitivity occur; and (b) the correlation of a just noticeable difference series with the corresponding energy values for such parts of the spectrum as are most frequently used. Neither of these possibilities, it is obvious, would be of much service for any very wide use of the eye as a measuring instrument. It will be the work of later papers to show in detail the selectiveness of the eye's response to wave-length and to intensity.

III. A CONVENIENT AND SENSITIVE RADIOMETRIC APPARATUS FOR WORK IN PSYCHOLOGICAL AND PHYSIOLOGICAL OPTICS.

The radiometric apparatus which we have used in our work for the past four years consists of two thermopiles, a very

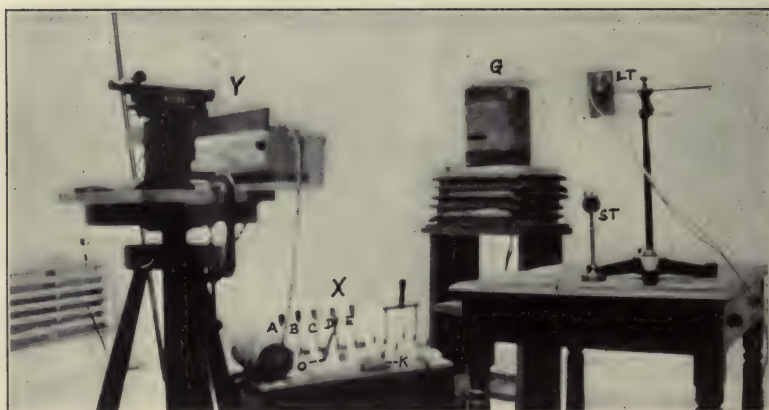


FIG. II

sensitive Thomson galvanometer and auxiliary apparatus for both thermopile and galvanometer. The two types of thermopile, the galvanometer, and the auxiliary apparatus for the thermopiles and galvanometer were constructed by Dr. W. W. Coblentz of the Radiometric Division of the Bureau of Standards. This apparatus is shown in Fig. II. LT is the linear thermopile in its brass mounting; ST is the surface thermopile; G is the galvanometer with its magnetic shields; X is the auxiliary apparatus for thermopile and galvanometer; and Y is the telescope and scale.

A. THE LINEAR THERMOPILE

We are at present using two types of thermopile, a surface and a linear pile. By using the two the energy of the light employed for the colored stimulus can be measured at three places: at the opening of the campimeter screen with the surface pile which has a receiving area just large enough to cover this opening; at the analyzing slit of the spectroscope with the linear pile; and at the eye also with the linear pile. The linear pile with a receiving surface of 2×12 mm. is broad enough to cover either the analyzing slit or the colored image of this slit in the plane tangent to the anterior surface of the eye. The linear pile measuring at the slit and at the eye would thus be adequate alone for our purpose. Some additional advantage is gained perhaps by having two instruments to serve as a check on each other and by measuring at three places instead of two.

The thermo-elements in both piles are of bismuth and silver. The linear pile consists of 20 elements joined in series. These elements are in the form of wire, the bismuth 0.1 mm. in diameter and the silver 0.051 mm. in diameter. The total resistance of the pile is 8.4 ohms; the area of the receiving surface is 2×12 mm. In Fig. III this thermopile is shown drawn to scale. In the lower right hand corner is shown a front view of the pile. The row of junctions in the center are the "hot" junctions or those exposed to the radiations to be measured. The rows on either side of this are the "cold" or unexposed junctions. Directly above is shown in detail the formation of a pair of "hot" and "cold" junctions. A bead of silver is used in welding the

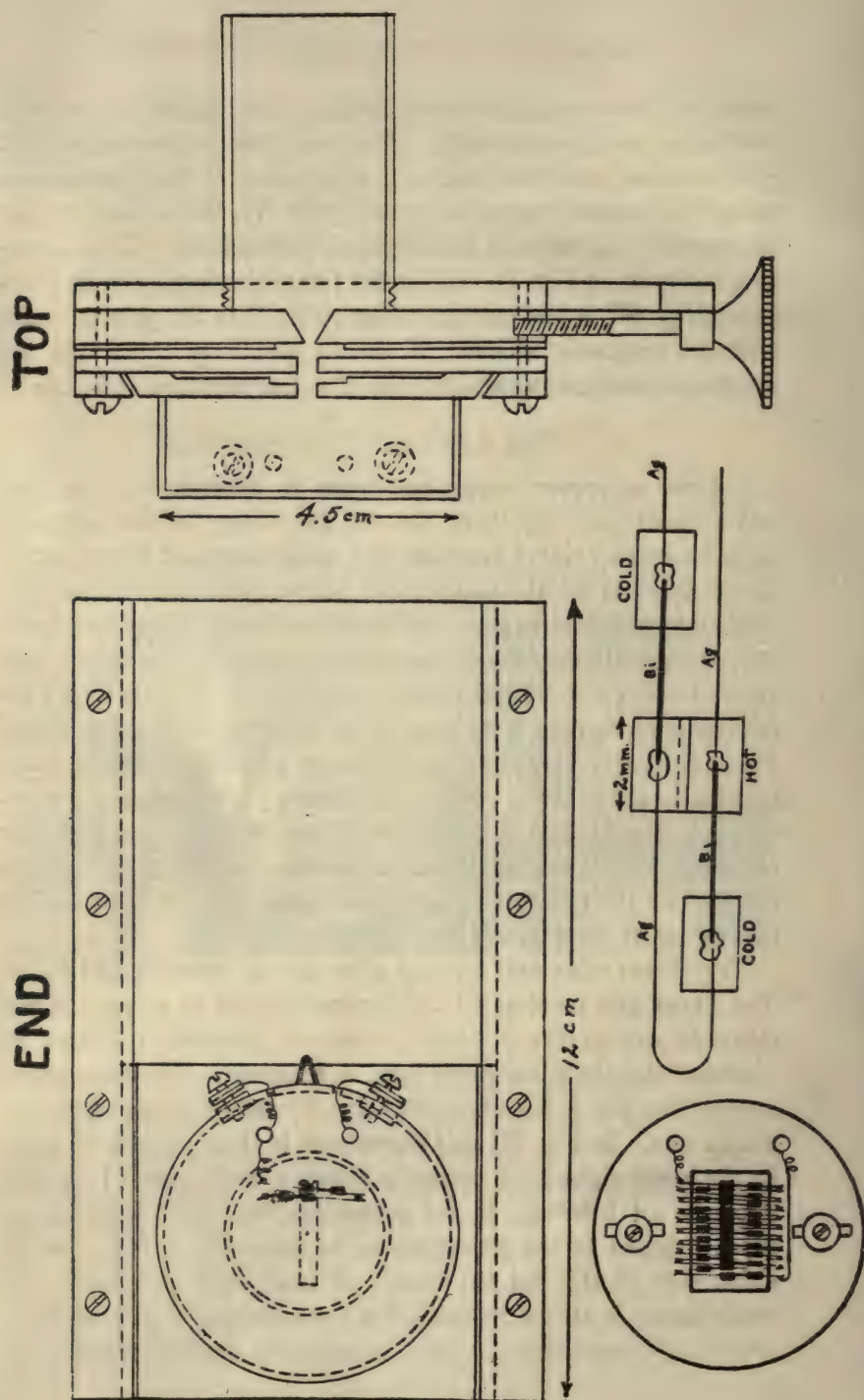


FIG. III

bismuth and silver wires together to form the junctions. Over each of the "hot" junctions is fastened a receiving surface of tin 2 mm. broad and long enough for the successive pieces to overlap. That the heat conducted to the "cold" junctions may rapidly radiate and thus maintain a temperature difference between the two junctions as large and as constant as possible, a surface of tin of appropriate size is fastened also over each of the "cold" junctions. In the diagram are also shown the top and end views of the pile mounted for use. The pile is mounted on a flat metal base which slides up and down in grooves constructed on the edges of the frame containing the analyzing slit for the spectroscope. When in use it is lowered so that the analyzing slit opens directly upon the face of the pile. During the color observation, when not in use, the pile is raised to the upper part of the frame clear of the slit, and is fastened by means of a small hook which engages the upper edge of the frame.

B. THE SURFACE THERMOPILE

This thermopile was designed especially for our work by Dr. Coblentz. The object was to get a thermopile that would measure directly all of the light that fell on the opening of the campimeter screen. This opening is 15 mm. in diameter. The surface of the thermopile was made 17 x 17 mm. The surface exposed to radiation was reduced to coincide with the stimulus-opening by means of a circular diaphragm 15 mm. in diameter. In order to shield the exposed junctions from the influence of air currents this aperture was covered with a thin sheet of clear glass (cover glass).

The pile consists of three units joined in parallel. Each unit consists of 20 elements, bismuth and silver wire, joined in series. The bismuth wire is 0.1 mm. and the silver wire is 0.051 mm. in diameter. The total internal resistance of the pile is 3.65 ohms. Each of the "hot" junctions is covered with a surface of tin, 6 x 1 x 0.02 mm. The "cold" junctions are to the rear of the "hot" junctions instead of on either side as is the case in the linear pile. That is, on leaving the "hot" junction, instead of running to either side each of the elements is bent backward. To

simplify the construction, the radiating surfaces of tin which cover the "cold" junctions in the linear pile were omitted from the surface pile.

A diagram of the surface pile is shown in Fig. IV. A shows the mounting for the ivory supports to which the sensitive elements are attached; B shows the way in which the wires are bent from the "hot" junctions to form the "cold" junctions; C gives

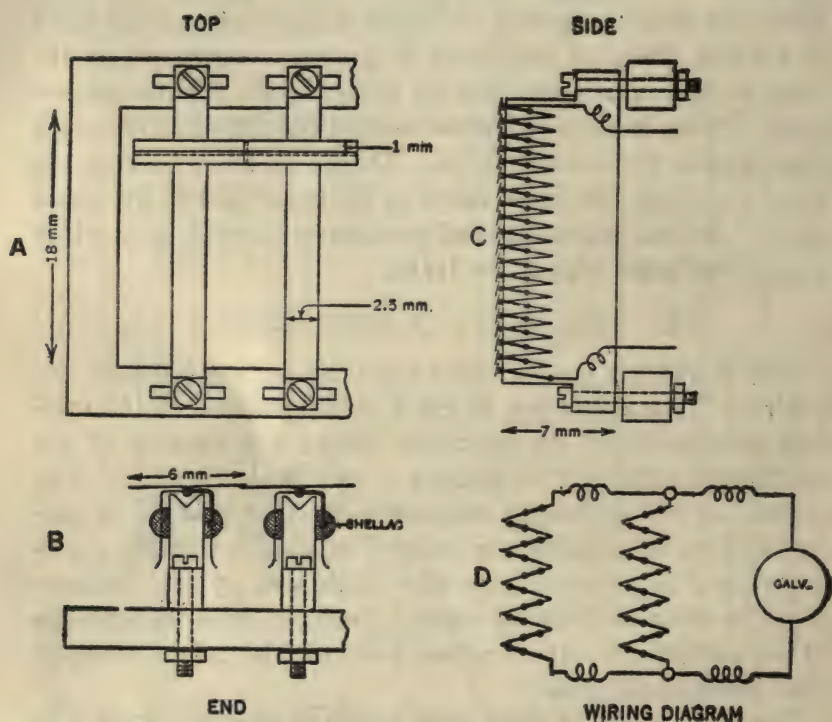
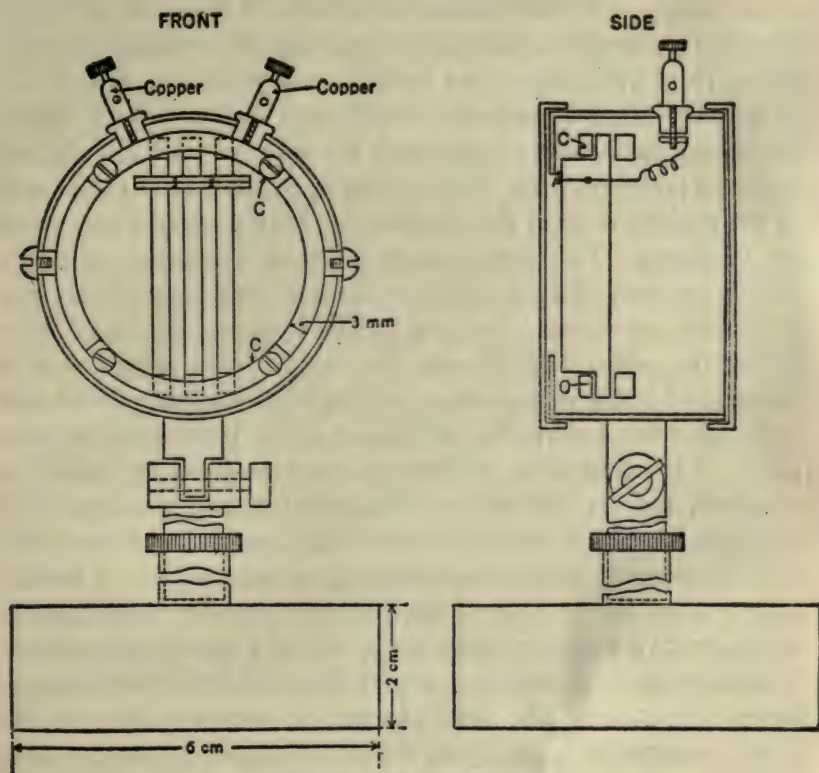


FIG. IV

a side view of one unit of the pile consisting of 20 thermoelements mounted on the ivory support; and D shows the way in which the units are connected. In order to reduce the resistance and thus increase the sensitivity of the pile they are joined in parallel. In Fig. V are shown the front and side view of the holder for the thermopile.



HOLDER FOR THERMOPILE

FIG. V

In common with all surface radiometers of high sensitivity this surface thermopile causes some drift of the zero of the galvanometer unless the instrument and the shutter enclosing it are carefully protected from changes in temperature due to contact with the air of the room. Winding the terminals and body of the thermopile with cotton batting has overcome this effect almost entirely.

C. THE RADIATION STANDARD

The thermopile might be used for equalizing energies or reproducing given light intensities without calibration. However, if its responses are to be converted into C. G. S. units, calibration

is necessary. For this a radiation standard is required. For this standard we employ a thoroughly seasoned carbon lamp the radiations from which have been carefully evaluated in terms of the primary standard conserved at the Bureau of Standards.¹ From the known radiations of this lamp the sensitivity of the pile per unit area is determined. This is done as follows. First the value of the radiations from the standard incident upon unit area must be computed. This computation is based ultimately upon the Stefan constant of total radiation from a black body: $T = 5.7 \times 10^{-12}$ watt per sq. cm. For our standard operated at 102.1 volts giving 0.4 ampere of current, the value of the radiation at a distance of 2 meters is 90.70×10^{-8} watt per sq. mm. With this radiation value known, the calibration of the thermopile becomes easy. It is set up at a distance of 2 meters from the radiation standard, and the deflections of the galvanometer, the sensitivity of which must have been determined just previous to the calibration, is obtained. From this value and the total amount of energy falling upon the surface of the pile, the amount of energy required to give a galvanometer deflection of 1 mm. is determined. This may be taken as a measure of the sensitivity of the radiometric apparatus. The total amount of energy falling on the pile is obtained by multiplying the radiation per sq. mm. at 2 m. by the area of the receiving surface of the pile and correcting this value for the absorption of the glass cover (12% in case of our instrument).

Since the galvanometer sensitivity may vary from time to time, it is necessary to establish for it a standard of sensitivity and to reckon the radiation sensitivity of the apparatus in terms of the reading at this standard sensitivity. In order to use the value so established in future work with the apparatus, it is necessary to determine each time the current sensitivity of the galvanometer which is quickly done with an especially devised testing apparatus, and to compute from this and the standard sensitivity a correction factor which has to be applied to all readings taken at this time.

¹ Coblentz, W. W. Bull. Bur. Standards, 1914, 11, p. 87.

D. THE GALVANOMETER

The galvanometer used was constructed specially for the thermopile employed. It is of the Paschen small-coil type, shielded from magnetic influence by four cylindrical soft iron shields. Its parts are as follows. The magnetic field of the instrument is given by four coils each having a resistance of 6 ohms. Each coil is wound in three layers, 2 ohms per layer. The wire used was B & S gauge Nos. 38, 30, and 26, single covered silk insulation, the diameter of the bare wires being respectively 0.101, 0.255, 0.405 mm., and the lengths 92, 595, and 1375 cm. Each coil has a diameter of 2.8 cm. and a thickness of 7 mm. The coils are joined in pairs, series parallel, giving a total resistance of 6.58 ohms. The needle system consists of two groups of six magnets placed above and below the mirror. The magnets are of tungsten steel and are from 1.5 to 2.5 mm. in length; from 0.3-0.4 mm. in width; and 0.1 mm. in thickness. They are mounted so that each group of magnets has the form of an ellipse. The mirror is of thin cover glass 2 mm. x 3 mm., platinized by cathode discharge. The magnet groups and the mirror are mounted on a segment of a very small glass rod in such a way that the centers of the groups are 33.5 mm. apart and the mirror is midway between them. At the lower end of the rod is attached a damping vane of bolometer platinum $5 \times 4 \times 0.003$ mm. The needle system weighing 12-15 mg. was made heavy to minimize the influence of earth tremors. It is suspended between the coils by means of an extremely fine quartz fiber.

The assembled galvanometer consisting of base provided with leveling screws, the coils and their ebonite supports, the needle system and the containing tube for its suspension, is drawn to scale in Fig. VI and needs little explanation. The coils, coated with paraffine to give the insulation needed, are attached to the ebonite supports by means of soft wax (a mixture of Venice turpentine and beeswax which hardens to the desired consistency on standing). Soft wax is used because it permits of an easy and convenient adjustment of the position of the coils, the faces of which should be brought to exact parallelism. This can

be done very conveniently by moving up and down between the faces of the coils a rectangular object of the proper width, *e.g.*, a microscope section glass. Having been rendered parallel the faces of the coil are brought into the vertical by means of the leveling screws on the base of the instrument. The quartz fiber which suspends the needle system passes through the containing tube and is attached at the top to a short brass pin with a milled head, held in place by means of a set screw. The needle system is fastened to the quartz fiber with shellac. The attachment of the needle system to the quartz fiber and this in turn to the brass pin is made on a special mounting board designed for the purpose. The housing of the galvanometer is of microscope section glass. This not only prevents disturbances of the needle system due to air currents but permits of an excellent illumination of the interior of the galvanometer.

The magnetic shielding²¹⁵ of the galvanometer consists of an inner laminated cylinder made up of six layers of transformer iron, $3\frac{1}{2}$ inches in diameter, and three sections of soft iron pipe 5, 6 and 10 inches in diameter respectively, and 10 inches high. Each of these shields contains a horizontal window 1 cm. high and 8 cm. long through which the mirror is viewed. They rest on a slab of slate and a glass plate is put on top. Care is taken to keep the shields well annealed, but in spite of this some magnetism is acquired in handling. The influence of this and of the earth's field on the needle system has to be overcome if the galvanometer is to have the long period needed for a high current sensitivity. This might be done in two ways. (1) The unknown and very complex field due to the iron shields could be ascertained by means of control magnets; or (2) a stronger and simpler known field could be created and this field be weakened to the desired amount. The latter is found to be the more feasible procedure. The field is created by a short magnet placed under the base of the galvanometer in such a position that the mirror and its reflected

²¹⁵ For working in localities close to street cars and other magnetic disturbances, it is necessary to use a more thoroughly shielded galvanometer, the coils of which are imbedded in soft iron. See Bull. Bur. Standards, 1916, 13, p. 423.

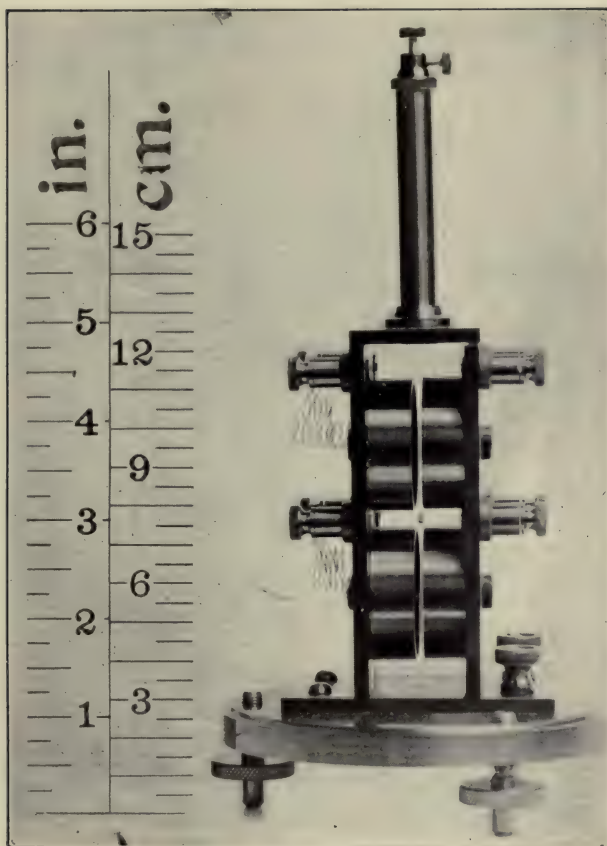


FIG. VI

scale are brought into the field of the observing telescope. This field is then weakened the desired amount by a larger magnet of greater strength placed on the glass plate covering the top of the shields. The large magnet is rotated about the vertical axis of the galvanometer until the needle which follows it passes quickly through a neutral point and rotates in the opposite direction. In this position of the magnets the effective field is not so great as the moment of torsion of the needle system, and the galvanometer should have a long period and a high current sensitivity.

E. THE AUXILIARY APPARATUS

This consists of a device for testing the sensitivity of the galvanometer, a set of resistance coils to cut down the throw of current from the thermopile, and a reading telescope and scale.

1. *The sensitivity tester for the galvanometer.* The sensitivity tester consists of a dry battery giving an E.M.F. of 1.43 volts, and of three shunt coils with the necessary switches mounted on a suitable base. The shunt coils have a resistance of 1000, 100 and 1000 ohms respectively and are designed to pass 1.43×10^{-8} ampere of current through the galvanometer. This divided by the number of scale divisions of the deflections produced gives the value of the current required to give a deflection of one scale division, or the sensitivity of the instrument. That is, the formula expressing the sensitivity is $i = \frac{1.43 \times 10^{-8}}{d}$, in which d is the

number of scale divisions in the deflection; 1.43×10^{-8} is the total strength of the current; and i is the amount of current required to produce a unit deflection. For example, the galvanometer we use when adjusted to a 3 second period single swing gives a deflection of 40 scale divisions. This gives a sensitivity of $\frac{1.43 \times 10^{-8}}{40}$, or

$$3.575 \times 10^{-10}.$$

The galvanometer may be connected either with the sensitivity tester or the thermopile by means of two-pole, double-throw knife switch, as is shown at x in Figure II. When connected with the shunt coils of the sensitivity tester, the circuit is closed with the

dry cell by means of a key shown at k. Sometimes when the galvanometer is adjusted to a high degree of sensitivity a big shift of the zero occurs when the double-throw switch is changed from the pole connecting the galvanometer with the sensitivity tester to the pole connecting it with the thermopile. This makes it necessary to readjust the magnets above the galvanometer to bring the zero again to the center of the scale. To avoid this a further means is provided for testing the sensitivity of the galvanometer without opening the switch which connects the thermopile with the galvanometer. This consists of an auxiliary coil of wire which is connected directly with the dry battery and a key to close the circuit, shown at o. This coil is mounted in a fixed position on the casing about the tube containing the suspension above the needle system. When a current is sent through the coil a deflection is given which is proportional to the current sensitivity of the galvanometer and in fixed ratio to the deflection that is produced when the same amount of current is sent through the shunt coils of the sensitivity tester to the galvanometer coils. This fixed ratio may be determined once and for all by closing each of these two circuits in turn with the dry cell and comparing the deflections produced.

2. *Special resistance coils.* The resistance coils designed to cut down the throw of current from the thermopile to the galvanometer are thrown into the thermopile circuit by means of a series of single-pole, single-throw knife switches shown at A, B, C, D, and E, Fig. I. When A is closed no resistance is added, the full current passes to the galvanometer, and the true deflection is produced. When, however, B, C, D, and E are closed, the current is made to pass through 10, 40, 100, and 191 ohms respectively, and the observed deflections must be multiplied by the following factors to give the true deflections: $B=1.657$; $C=3.63$; $D=7.57$; and $E=13.55$.

3. *The telescope and scale.* The telescope and scale used are of the kind ordinarily employed when precision is wanted in the reading of a sensitive galvanometer. The scale is graduated to millimeters and illuminated by two 40-watt tungsten lamps which can be moved along in front of the scale to give the maximum

illumination of the part of the scale reflected by the mirror. The telescope is fitted with a lens system that will permit the clear reading of the scale at a distance of 2 meters. At this distance the definition is such that the scale can be read to 0.5 mm.

We recommend to workers in psychological optics the apparatus described in the foregoing pages as feasible, adequately sensitive, and precise. We feel that especial acknowledgment is due to Dr. Coblenz for a notable contribution to the apparatus available for the more definitely quantitative work in the subject.

Vol. XXIV
No. 3

PSYCHOLOGICAL REVIEW PUBLICATIONS

Whole No. 104
1917

Psychological Monographs

EDITED BY

JAMES ROWLAND ANGELL, UNIVERSITY OF CHICAGO

HOWARD C. WARREN, PRINCETON UNIVERSITY (*Review*)

JOHN B. WATSON, JOHNS HOPKINS UNIVERSITY (*J. of Exp. Psych.*)

SHEPHERD I. FRANZ, GOVT. HOSP. FOR INSANE (*Bulletin*) and

MADISON BENTLEY, UNIVERSITY OF ILLINOIS (*Index*)

STUDIES FROM THE PSYCHOLOGICAL LABORATORY OF THE UNIVERSITY OF CHICAGO

Transfer of Training and Retroaction A Comparative Study

By

LOUIE WINFIELD WEBB, Ph.D.

Instructor in Psychology and Education,
Northwestern University

PSYCHOLOGICAL REVIEW COMPANY

PRINCETON, N. J.

AND LANCASTER, PA.

AGENTS: G. E. STECHERT & CO., LONDON (2 Star Yard, Carey St., W. C.);
LEIPZIG (Koenigstr., 10); PARIS (76, rue de Rennes)



ACKNOWLEDGMENT

My main obligation for advice and suggestion in this experiment is due to Professor Harvey A. Carr. I owe much to Professor James R. Angell for encouragement and inspiration. I wish to express my sincere thanks to the many subjects who gave of their time and patience throughout the course of the experimentation.

CONTENTS

- I. INTRODUCTION.
- II. TRANSFER OF TRAINING.
 - A. Dependence of Transfer upon the Character of the Second Problem.
 - B. Dependence of Transfer upon the Character of the First Problem.
 - C. Dependence of Amount Saved upon Direction of Transfer.
 - D. Locus of the Transfer in the Learning Curve.
 - E. Selective Effect of Transfer upon Types of Error.
 - F. Summary.
 - G. Theoretical Discussion.
- III. RETROACTION.
 - A. Retroactive Effect of a Certain Activity upon Various Other Activities.
 - B. Retroactive Effect of Various Activities upon the Same Process.
 - C. Summary.
 - D. Theoretical Discussion.

TRANSFER OF TRAINING AND RETROACTION

I. INTRODUCTION

This study deals with the problems of transfer of training, retroactive inhibition, and their possible interrelations. The maze activities were utilized as the material of study, and both human and animal subjects were employed in the experimentation.

The term 'transfer of training' has long been in use in educational literature. The question usually considered in reference to this idea is, whether or not the learning of one problem aids, hinders, or has no effect upon the acquisition of a second problem. We shall employ additional terms to designate the three possible effects above mentioned. The term 'positive transfer' will be used to denote the results obtained when the learning in the first situation aids the learning in the second situation. 'Negative transfer' will be the term used to denote the results obtained, when the learning in the first situation hinders the mastery of the second problem. There is also the possibility that the first learned material will have no effect upon the acquisition of the new material. In this case transfer of training will not be present. Results obtained in this last named instance will be referred to under the designation 'absence of transfer.'

The term 'retroactive inhibition' has been definitely recognized since the experiments of Müller and Pilzecker,¹ and this term refers to the disturbing effect that the learning of a second problem has upon the retention of material previously acquired. As with transfer, there are three possible retroactive effects. The second acquired material may aid, hinder, or have no effect upon the retention of the first learned problem. We shall use the term 'retroaction' to refer to any possible effect that a second activity may have upon the retention of a first problem. The term 'posi-

¹ Müller and Pilzecker. Experimentelle Beiträge zur Lehre vom Gedächtniss. Ztsch. f. Psychol., 1900, Ergzbsbd. I.

tive retroaction' will be used to refer to the results when the second learned problem aids in the retention of the first problem. When the second problem interferes with the retention of the problem previously acquired, this result will be termed 'negative retroaction'; this last described condition is what Müller and Pilzecker termed retroactive inhibition. There is the further possibility, as was found by DeCamp,² that the second learned problem may have no effect upon the retention of the first problem. Such a case will be referred to by the term 'absence of retroaction.'

We believe that all three results have been obtained in the experiments on transfer. A positive or beneficial effect has been the usual result, and practically all the writers agree on the term positive transfer to denote this result. The term negative transfer seems to be employed to refer sometimes to the absence of any effect and sometimes to an inhibitive or detrimental effect. Terms are needed to discriminate between these two cases. Experiments on retroaction report but two types of results, either an interference or a lack of any effect. However, the possibility of all three effects must be recognized, and distinct terms are needed. Those suggested by this paper serve this need, and give the added advantage of a uniform terminology in comparing our two problems.

The maze was chosen as the basis of these experiments for three general reasons. (1) To study the phenomena of transfer and retroaction in a type of human activity somewhat different from those previously used, viz., an activity of a sensori-motor and adaptive character. (2) To investigate these phenomena in the animal as well as in the human realm. (3) To determine the essential similarity or difference between human and animal organization for these aspects of the learning process. The value of the maze problem for these purposes is obvious.

1. The experiments with human subjects on the problem of transfer have been rather adequately reviewed by Thorndike,³

² DeCamp: A Study of Retroactive Inhibition. *Psych. Rev. Mon. Sup.* vol. 19.

³ Thorndike: *Educational Psychology*, Vol. 2.

and Coover.⁴ There is no intention here of repeating this type of work. The object we have in view is simply to present a sufficient review of the literature to contrast the materials and methods previously used with those employed in this experiment. Our purpose is well served by following Coover's analysis. He classifies the experiments on transfer under the following topics. (1) Habituation to distraction. Vogt's work on testing the effect of reacting to a metronome, and reciting series of letters, upon the simultaneous activity of adding a column of figures serves as an example of this type of experiment. (2) Sensitivity. An illustration of this class is afforded by the work of Epstein, who sought to determine the influence of sound stimulations upon the acuity of simultaneous visual processes. (3) Discrimination. Bennett tested the effect of training in discrimination of shades of blue upon the discrimination of shades formed by a mixture of two colors, and upon discrimination of pitch differences. (4) Association. Thorndike and Woodworth tested the effect of training in estimating areas, weights and lengths upon estimations of various other areas, lengths and weights, of a different character. (5) Reaction. Thorndike investigated the effect of training in marking out words containing both the letters *e* and *s* upon the ability to mark out words with different pairs of letters, such as *e-g*, *i-t*, *s-p*, *e-r*; the length of the lines, size of type, and style of reading matter varied in the test series from that which was employed for the same items in the training series. The effect of the above mentioned training was also tested upon marking out misspelled words and words containing the capital letter *A*. Bergström reports some experiments on discriminative reactions, which showed interference. He tested the effects of training in the sorting of cards by one method upon the ability to assort cards by a different method. (6) Memory. The work of James in memorizing poetry, Ebbinghaus' work on memorizing nonsense syllables, the experiments of Ebert and Meumann with nonsense and other material, serve as illustrations of the work done in this field. (7) Voluntary control.

⁴ Coover: Formal Discipline from the Standpoint of Experimental Psychology. Psych. Rev. Mon. Sup. Vol. 20.

Judd and Cowling tested the effects of drawing an imaged form with the eyes open upon the ability to draw the image with the eyes closed.

We wish to offer here criticism of one aspect of Coover's analysis. We are inclined to believe that his classification is too broad. We doubt the advisability of including under the term transfer of training his first and second headings. These have to do with the effect that one activity has upon a simultaneous activity. Undoubtedly some sort of interaction does obtain between two such activities. Transfer of *training*, however, implies the utilization of the effects of training or a learning process upon some subsequent activity. Coover's two classes deviate from this definition in two respects. The first activity hardly involves anything that may legitimately be termed training or learning, and the effect is upon a simultaneous and not upon a subsequent process. Coover, however, discusses these experiments under the topic of 'Formal Discipline,' but to our mind, the two terms formal discipline and transfer of training, have been used as synonyms in the literature.

In further considering this classification, it appears to us that three divisions, which Coover does not mention, may be added. (8) Attention and Reproduction. Coover used in his experiments some of the methods and materials mentioned under the above headings, such as marking out words, estimating weights, and discrimination problems. He also performed experiments with other materials that we would classify under another heading. In the experiments on attention Coover tested the effects of training in activities which involve a large amount of attention such as, tachistoscopic work, learning 12-letter-rectangles, reactions to sounds, and memory training, upon such activities as reactions to sounds, marking out a's and o's, card sorting, memory of visual signs, trains of ideas, tapping, and many other activities. The difference in attention was noted between the training work and the testing series. In the reproduction experiments the effect of training in sound discrimination was tested upon such activities as recognition of one or two letters, reproduction of letters, sound discrimination, and memory for visual

symbols. (9) Cross education. The experiments on cross education note the transfer to one hand or foot from a training of the other hand or foot. The experiments of Davis,⁵ and Johnson⁶ with tapping; those of Woodworth⁷ on hitting a dot with a pencil; the ones of Swift on tossing balls; and those of Starch on tracing an outline seen only in a mirror are cited as illustrations of work in cross education. (10) Sensori-motor learning of an adaptive character. At the meeting of the American Psychological Association, 1915, Dr. Frank N. Freeman reported some tests made by him on transfer in sensori-motor learning. This is a study of mirror drawing. The apparatus was so constructed that it was modifiable in such a way as to vary the conditions indefinitely. The problem in these experiments was to test the effect of learning to connect six dots with a pen, with the mirror in one position, upon learning to connect the six dots with the mirror in another position.

There are marked variations in the results from the various experiments on transfer of training. In summarizing the experiments, following his review, Coover interprets the results in most instances as evidence of positive transfer; he recognizes, however, several instances of negative transfer. Thorndike, in his review of the literature, gives the reader a strong impression that the evidence for positive transfer is rather weak. This is indicated by the following quotation. "These experimental facts as a whole, like those concerning memory, leave a rather confused impression on one's mind, and resist organization into any simple statement of how far the improvement wrought by special practice spreads beyond the function primarily exercised."⁸ We are of the opinion that some of the experiments demonstrate the existence of a decided positive transfer. Bennett showed that training in memorizing poetry improved memory for digits and names of places; Ebert and Meumann proved that training for nonsense syllables improved the memory for letters, numbers, words, poetry, prose and optical symbols. Some of the results

⁵ Davis: *Researches in Cross-Education*. Yale Studies, Vol. 4.

⁶ Johnson: *Researches in Practice and Habit*. Yale Studies, Vol. 4.

⁷ Cited by Thorndike. *Op. Cit.* p. 366.

⁸ *Op. Cit.* p. 416.

also demonstrate the existence of negative transfer. A splendid illustration of this is found in the work of Bergström, wherein he showed that the training in the sorting of cards by one method interfered with the sorting of cards by a new method, and that in learning series of nonsense syllables successively, the time becomes progressively longer if the series possess recurring elements. Further, we believe that some of the results indicate an absence of transfer. An illustration of this is taken from the work of Sleight as reviewed by Thorndike⁹: Sleight's results show that training in memorizing poetry has very little, if any, effect upon the memorizing of tables of figures or prose. James interpreted his own experiments on memory as evidence of an absence of transfer. Some of Thorndike and Woodworth's experiments give evidence of little or no transfer.

We have seen that the materials employed in the experiments on transfer vary from those of a purely ideational character, such as memorizing poetry or nonsense syllables, to those having less and less of an ideational character; in some of the experiments, such as those dealing with visual and auditory discriminations, the sensory elements predominate; there are also experiments dealing mainly with material of a sensori-motor character such as those of Freeman referred to above. The material used in our experiment is more nearly like that employed by Freeman. The maze activities are essentially sensori-motor in character. The sensori-motor element predominates in mastering a maze situation. This is the universal conclusion of all experimenters with such animals as the white rat. The rational element is present with human subjects, but it is recognized that these ideational activities function with little effectiveness in the mastery of the problem. The sensory elements employed are what James calls resident sensations. Under the conditions of our experiment the effective sensory factors are tactual and kinaesthetic in character.

There have been no published experiments dealing specifically with transfer in the animal field; neither has the maze activity been utilized in the study of this phenomenon with human sub-

⁹ Op. cit., p. 379.

jects. Perrin¹⁰ used the same subjects in the learning of the several mazes in his experiments, but he made no definite report of the transfer effect.

There have been no experiments on retroaction which employed sensori-motor activities. Only three important studies with human subjects have been reported. Müller and Pilzecker¹¹ investigated the retroactive effect of memorizing nonsense syllables and observation of pictures upon the retention of previously learned material. They reported that such subsequent activity exerted a negative or inhibitive effect upon the retention of the first learned series. DeCamp¹² investigated the retroactive effect of such activities as memorizing nonsense material, multiplying a series of numbers, the solution of various problems, and playing chess, upon the retention of previously mastered material, and he decided that no retroactive effect was present in his experiments. Heine¹³ repeated Müller and Pilzecker's experiments and verified the former results; she also discovered negative retroaction between the letters of the syllables; she made further tests for retroaction employing recognition instead of learning, and here she failed to secure evidence to prove the presence of a retroactive effect.

It will be noted that Müller and Pilzecker investigated the retroactive effect of two similar activities of an ideational character. DeCamp also used activities of an ideational character, and employed dissimilar as well as similar activities. We shall also use two similar activities in our experiments—the effect of one maze activity upon another—but shall employ sensori-motor activities rather than those of an ideational character. This difference of material is significant because of a possible difference in the stability of the two types of activity. Results obtained from stable material may thus indicate that the existence or degree of negative retroaction may depend in part upon the

¹⁰ Perrin: *An Experimental and Introspective Study of the Human Learning Process in the Maze*. Psych. Rev. Mon. Sup. Vol. 16.

¹¹ Op. cit.

¹² Op. cit.

¹³ Heine: *Über Wiedererkennen und rückwirkende Hemmung*. Ztsch. f. Psychol., 1914, Band 68.

stable character of the activities employed in the experimentation.

The results of DeCamp suggest the further possibility that retroaction may be in part dependent upon the similarity of the two activities employed. For this reason we decided that it would be better to utilize two highly similar processes. Our method deviates from that of DeCamp, and Müller and Pilzecker in two other important respects: (1) The temporal relation of the two activities. Their first activity is learned with continuous trials in a single sitting, while the second work is performed within five to fifteen minutes subsequently. Rats of necessity master the two mazes with a distribution of trials extending over several weeks, with at least a day's interval between the completion of the first maze and their introduction to the second. The same distribution of effort was maintained for the human subjects in order to secure comparable results in the two cases. (2) The method of measuring retention. The previous experiments measured the amount retained after the interpolation of the second activity by the usual method of verbal reproduction. Such a method is impossible with the maze problem. In our experiment retention was necessarily measured by the relearning method.

2. Our second purpose is justified by the fact that there has been little or no systematic work done on either problem in the animal field.

Many of the early workers in the animal field incidentally noted and commented upon the presence of transfer when animals used on one problem were employed on some other task.¹⁴ There are but three systematic works on this subject with which the author is acquainted. One was reported by Mr. H. H. Wylie at the meeting of the American Psychological Association, 1915. He trained white rats to respond to a certain stimulus, e. g. light or pain, in a given situation. After this habit was well established, the animals were taught to respond to a different stimulus in the same situation, and the degree of transfer between these

¹⁴ Watson has given an excellent review of the literature in this field in his text, "Behavior, An Introduction to Comparative Psychology." Chapter VI.

two highly similar situations was measured. Hunter¹⁵ demonstrated with white rats that the acquisition of an auditory-motor habit interfered in the formation of a new habit of opposite character. Pearce¹⁶ has demonstrated the same result, duplicating the conditions of Hunter, but employing visual-motor habits.

Within the knowledge of the author of this paper, no experiments on retroaction in the animal field have been reported.

3. We also wished to determine whether the laws and conditions of transfer and retroaction are the same for both humans and animals; whether human and animal organizations are essentially identical in kind.

Thorndike,¹⁷ in elaborating the laws of learning, maintains the validity of the above proposition. Most Comparative Psychologists would subscribe to this belief, but it is a belief or generalization which represents to a large degree a working hypothesis, the testing of which Comparative Psychology recognizes as one of its main tasks. Thus far Comparative Psychology has demonstrated the hypothesis in many respects, e. g. both humans and animals can learn, and both can learn by the same method. However, the proposition is not completely demonstrated in every respect, and we wish in these experiments to test the hypothesis in two additional phases, transfer of training and retroaction.

To test the proposition that the laws of learning are the same for humans and animals, it is advisable to secure comparisons where the problems and conditions are as similar as possible. Many more comparative statements and generalizations could be made from the work already done, if similar situations had been used. So far as the author is aware, there have been only three experiments reported, wherein the identity of the problems and situations was maintained to an important degree. Hicks and Carr¹⁸ compared the ability of white rats and humans in

¹⁵ Hunter: The Interference of Auditory Habits in the White Rat. *Jour. Animal Behav.*, Vol. 7, No. 1.

¹⁶ Pearce: A Note on the Interference of Visual Habits in the White Rat. *Jour. Animal Behav.*, Vol. 7, No. 3.

¹⁷ *Op. Cit.*, p. 12f.

¹⁸ Hicks and Carr: Human Reactions in a Maze. *J. Animal Behav.*, Vol. 2.

learning the maze problem. The two problems were similar in kind, but the mazes were not identical in pattern, and some other conditions differed so that their situations were not as similar as they might have been. Hunter¹⁹ was the first one to study humans and animals with conditions and problems identical. He tested animals and children on the problem of delayed reaction. The identity of the two situations was maintained as far as possible, and he found that the period of delay in the reaction could be lengthened as he ascended the animal scale from rodents, dogs, primates, and children. Pechstein²⁰ has studied various methods of motor learning, using animals and humans in learning the maze. His maze patterns, his technique and method of procedure were similar to a marked degree. Our experiment is only one more step of the many that are needed in the comparative field to test the above hypothesis. The effort was made to have the situations as comparable as possible, and it is believed that another bit of evidence is added by this study towards the solution of this important problem.

In the experiments with the rats two adjustable mazes were used. They were the same in size, 4' x 3'8" x 6". These mazes were made of $\frac{1}{2}$ " stuff and were covered with glass. Each maze was placed upon a frame about eighteen inches high. The runways and the cul de sacs were four inches in width. The partitions within the mazes were made of galvanized sheet iron, and were set in brass supports, so that they were easily moved back and forth. This construction gives large possibilities of altering the location of the cul de sacs in relation to the true path. Six different maze patterns were thus constructed and utilized in the experiments with rats. These patterns will be hereafter designated as Mazes A, B, C, D, E, F. They are represented in Figures 1 to 6.

The mazes used for the human experiments are known as 'pencil mazes.' Two of these were made from solid aluminum

¹⁹ Hunter: The Delayed Reaction in Animals and Children. *Behav. Mon. Vol. 2.*

²⁰ Pechstein: Whole vs. Part Methods in Motor Learning. *Psych. Rev. Mon. Sup. No. 99, Vol. 23.*

castings and two from solid brass castings. The cul de sacs and the true runways were milled out of the castings, and were $\frac{1}{4}$ " wide by $\frac{1}{4}$ " deep. The partitions were $\frac{1}{4}$ " wide, and the outside dimensions were $5\frac{3}{4}$ " x $5\frac{1}{4}$ ". But four patterns were employed with human subjects, and these are designated as pencil mazes A, B, C, D. These four duplicate exactly, except on a reduced scale, the patterns of mazes A, B, C, and D respectively used with the rats. They were identical as to the location of runways and the cul de sacs. Each section of the true path, each turn in the true path, and each cul de sac had the same relative position in each type of maze. Not only did we have identity of maze patterns, but we endeavored to have the technique and method of experimentation, which will be described later, as similar as possible. The exact duplications of maze patterns and the identity of method of procedure for rat and human subjects were adopted in order to achieve our third object mentioned above, the comparison of human and animal results upon identical problems.

The rats used in this experiment were thoroughly tame, having been fed for a week in the maze before the experimentation began. The animals were kept in the same room where the experiment was performed, and the location and conditions of their living cage were not changed during the experimentation. When the cages were cleaned, this work was always done after the experiment of the day was finished, thus giving the rats twenty-four hours to become adjusted to the disturbance. The work of each day was done in the late afternoon. Before the experimentation began the window shades were lowered and the electric lights were switched on, thus giving uniform lighting conditions throughout the experiment. The animals were run under a stimulus of normal hunger. Each animal was given one trial per day for the first four days and two trials per day thereafter until the maze was mastered. At the end of the first trial the animal was allowed a bite or two of food, and then immediately was given his second trial. When the group had finished the day's work, they were allowed to eat for seven minutes of a diet of bread and milk; at intervals sunflower seed was added to the

diet. In the experiment on transfer, after each subject had mastered the first maze, he was transferred to the second maze on the following day. The same method of procedure obtaining in the learning of the first maze was continued during the mastery of the second maze. During the period in which the rats were not running the maze, they were each day taken from the cage and allowed to run on the top of a table for exercise. They were fed as during the experiment outside of the cage. This was done in order to maintain the normal conditions of the experiment and to obviate any disturbance until after the tests for retention were given. The results presented in this paper were obtained from one hundred and thirty-six white rats. The animals were from seven to twelve weeks old at the beginning of the experiment; they were in good condition and remained so throughout the tests. Males and females were in each of the groups in the separate parts of the experiment.

The human subjects upon entering the room in which the experiment was conducted, were seated at one side of a table. On this table the pencil maze was placed. Light strips of wood were nailed to the table at each side of the maze to prevent any movement when a subject was working on it. On the table, and covering the maze, was a frame one foot high on one side and one and one half feet high on the other side. The frame was eighteen inches wide and sixteen inches deep, and was covered with a heavy black cloth which hung loosely on the side towards the subject, and adequately hid the maze from his sight. The higher end of the frame opposite the subject was uncovered. This arrangement left the maze in full view of the experimenter, so that all errors were easily noted. This method was adopted in order to eliminate the disconcerting effect of a blind-fold, and to secure a comparable situation as to vision for humans and animals. The human subjects traced the maze with a stylus made of hard rubber. A shoulder about one inch from the lower end prevented the hand from slipping and coming in contact with the maze. The lower end of the stylus was $3/16''$ in diameter and the pathway, being $1/4''$ wide, permitted easy contact with both sides. Each subject was given the following instruc-

tions: "Please put your hand under the cover; grasp the stylus and hold it as erect as possible with comfort; keep the stylus in the groove, and explore the area assigned until you are told to stop. Use any method you desire and think as much about the problem as you wish during the experiment, but do not try to draw the area and try not to think about the problem during the interval between successive trials." At the end of the run there was an opening in the pencil mazes, corresponding to the food box in the other mazes. This enabled the subjects to know when the end was reached, and thus they had to be told when to stop working for only one or two trials. One trial per day was given to each subject for the first four days and two trials per day thereafter until the problem was mastered. Sunday necessarily had to be omitted with the human subjects. When each subject had mastered their first maze, the following day he was transferred to the second maze. None of the subjects knew that the experiment was dealing with the problem of transfer and retroaction, and in no instance was a subject told that the maze had been changed; he discovered the new situation by going through it and noting the disturbance. A large part of the subjects were naïve, and knew nothing of the maze problem. Here again we notice the similarity of the procedure for humans and animals; each had to discover the new problem empirically. When a human subject finished the first part of the experiment, he was asked to come back at some later specified time for another part of the experiment. Each subject was requested not to think about the problem any more than possible, nor to try to draw the maze pattern during the interval. In the test for retention, the same instructions were given, and the same conditions were maintained as were used in the first part of the experiment. At no time was a subject told that this other part of the experiment was to be a test for retention. The subjects, fifty-two in number, were graduate and undergraduate students, all of whom were studying psychology in the University of Chicago. There were both men and women in each of the groups that were used in the various parts of the experiment.

Individual time and error data for each trial were recorded.

The time in seconds was measured with a stop watch. The time for each trial was counted from the moment the subject entered the maze until he reached the entrance to the food box. No record of the distance traversed was kept, as this was practically impossible under the conditions here maintained. We agree with Mrs. Hicks²¹ in her emphasis upon the distance traversed, as a criterion for measuring the learning process in a maze situation. However, our method of counting errors overcomes, to a large extent, this deficiency; each section of the true path, no matter how short, was counted an error in retracing. The unit of error was one section of the pathway. Each entrance into a cul de sac, whether for all or part of its length, each retrace over the whole or a part of a section of the true pathway, was counted an error. The errors due to entering a cul de sac while going forward, the errors due to entering a cul de sac while returning home, and the errors due to retracing the true pathway, were kept separately. The criterion of mastery was four perfect trials out of five, and in these five trials not more than two errors were allowed.

The comparableness of the technique and methods of procedure for both types of subjects is obvious. The maze patterns are identical, the mazes for the humans being duplicates of those used with the rats, with the exception of a difference in size. The methods of recording the data were the same; the distribution of effort is practically the same. Watson²² has shown that rats make very little use of vision in mastering the maze problem; the rats have eyes, but do not use them in this situation. With the technique here maintained, it can be said that the humans have eyes but do not use them in this situation, and further the disconcerting effect of a blind-fold is absent. Both humans and animals come to the problem naïvely. The humans were not informed as to the situation; they had to discover their problem as did the rats. This last condition was more fully maintained

²¹ Hicks: *The Relative Value of the Different Curves of Learning*. Jour. Animal Behav., Vol. 1.

²² Watson: *Kinaesthetic and Organic Sensations*. Psych. Rev. Mon. Sup. Vol. 8.

with the human subjects who learned pencil mazes C and D. These subjects knew nothing of the problem until after they worked on it. The groups that learned pencil mazes A and B, involving transfer, were graduate students in Psychology and were well enough acquainted with the experiments being carried on in the laboratory, to know that this experiment was dealing with the maze problem. The results of the naïve and sophisticated groups were compared, and only minor differences, such as can be accounted for in terms of chance differences, were noticeable. However, all of the human subjects came to the transfer and the tests for retention naïvely, as did the rats, as nothing was said to any of the subjects to indicate the changes. During the interval that the rats were not running the maze, the conditions of exercise and feeding were not changed, so normal conditions obtained when they returned to the maze for testing retention. The human subjects went about their normal activity, and were asked not to think about the problem, nor to try to draw the area during the interval. Thus it is observed to what extent we endeavored to keep the conditions similar for the two types of subjects. By these means our third purpose mentioned above is well served, and our conclusions in comparing the transfer and the retroactive effects on humans and rats are rendered more valid.

II. TRANSFER OF TRAINING

A. DEPENDENCE OF TRANSFER UPON THE CHARACTER OF THE SECOND PROBLEM.

The object of the first experiment is to determine to what extent the nature and degree of transfer is a function of the second problem.

In educational literature there has been much discussion relative to the question of the general spread of training. The classical argument has been that the study of Latin or Greek improves the intelligence of the student in such a manner that said learner will be able to acquire more easily any other subject he thereafter studies. This statement seems to imply that the existence of any transfer effect in subsequent situations depends wholly upon the character of the previous training. The other side of the question must also be recognized, viz. that the functional efficacy of classical training may depend in large part upon the nature of one's subsequent mental activity—that it may have more effect in the life activities of a lawyer than of an electrical engineer. Our first experiment was designed to solve this latter question, not that we experimented with the Classics and other strictly intellectual subjects, but rather that we were concerned with the solution of the problem in the realm of motor learning.

In terms of mazes the problem is readily illustrated. Several groups of subjects first learn a common maze A, and each group subsequently learns a different maze. One group is thus transferred from Maze A to Maze B, another from A to C, one from A to D, another from A to E, and one from A to F. If the nature and degree of the transfer is wholly a function of the first maze, fairly uniform results should be secured for all groups. The group differences should be only such as can be attributed to chance or group factors. Marked variations in the results would indicate, on the other hand, that the nature and degree of transfer is in part a function of the nature of the second maze.

In our experiment six groups of subjects first learned Maze A. These groups comprised a total of 54 rats and 21 humans. A group of nine rats and a group of five humans were then transferred to Maze B; a group of eleven rats and a group of six humans subsequently learned Maze C; a group of six rats and one of five humans were transferred to Maze D; a group of eight rats then mastered Maze E, and a group of nine rats subsequently learned Maze F.

The transfer effect is measured by the difference between 'original learning' and 'transferred learning.' By original learning we mean the acquisition of a maze by a group of subjects without previous maze experience. By transferred learning we refer to the mastery of a maze by a group with a previous maze experience. Thus control groups were necessary for Mazes B, C, D, E, F, in order to secure data on the original learning. These control groups consisted of 20 rats and 10 humans for Maze B, 11 rats and 10 humans for Maze C, and 13 rats and 11 humans for Maze D. No human subjects were employed on Mazes E and F, and the control groups for these mazes consisted of 16 and 15 rats respectively.

Table 1 presents the results in the absolute terms of averages for our first experiment. The transfer effect was measured separately in terms of trials, errors and time, thus giving us three criteria of measurement. The group averages, together with the average deviation, for the original and the transferred learning are given for each of the criteria. The letters in the table indicate the records for the several mazes; thus *B* indicates the record for the original learning of that maze, and *A-B* denotes the transferred learning of the same maze. The *A* in connection with the *A-B* means that Maze A constituted the previous maze experience of the group. The symbol *Sav.* refers to the average amount saved for each group and these figures thus measure the transfer effect for each group. Table 2 states the saving in relative or percentage terms. These percentages of transfer are secured by dividing the absolute amount saved for any maze by the figure measuring the original mastery of that maze. If 12 trials are saved in transferring to Maze B, while 15 trials were

TABLE 1. COMPARATIVE RECORDS OF ORIGINAL AND TRANSFERRED LEARNING.

RATS			
	Trials	Errors	Time
<i>B</i>	56.2 ± 14.7	$224. \pm 71.8$	2468.6 ± 1614.3
<i>A—B</i>	$12.9 \pm 9.$	31.8 ± 20.3	400.5 ± 200.2
Sav.	43.3	192.2	2068.1
<i>C</i>	45.7 ± 14.5	238.5 ± 76.9	2939.4 ± 2504.5
<i>A—C</i>	18.3 ± 14.2	129.3 ± 78.3	1902.2 ± 1261.8
Sav.	27.4	109.2	1037.2
<i>D</i>	16.7 ± 8.1	$153.6 \pm 51.$	2777.3 ± 1272.4
<i>A—D</i>	5.2 ± 4.4	31.2 ± 20.6	487.1 ± 218.6
Sav.	11.5	122.4	2511.1
<i>E</i>	4.4 ± 2.5	$19. \pm 6.$	213.5 ± 75.5
<i>A—E</i>	3.5 ± 1.5	8.6 ± 5.9	78.1 ± 53.6
Sav.	.9	10.4	135.4
<i>F</i>	27.9 ± 10.1	128.5 ± 39.2	1209.8 ± 447.7
<i>A—F</i>	10.3 ± 6.1	$73.8 \pm 37.$	490.7 ± 294.7
Sav.	17.6	54.7	719.1
HUMANS			
<i>B</i>	33.6 ± 14.3	285.2 ± 205.4	$1166. \pm 514.2$
<i>A—B</i>	10.8 ± 5.9	32.4 ± 13.7	149.4 ± 54.7
Sav.	22.8	252.8	1016.6
<i>C</i>	$38. \pm 9.8$	133.7 ± 43.8	736.4 ± 124.2
<i>A—C</i>	$30.5 \pm 13.$	106.8 ± 44.5	521.5 ± 168.1
Sav.	7.5	26.9	214.9
<i>D</i>	11.6 ± 6.9	$203. \pm 176.$	$773.5 \pm 498.$
<i>A—D</i>	5.6 ± 3.3	$11. \pm 10.$	87.2 ± 64.6
Sav.	6.	192.	686.3

expended in its original mastery, it is evident that the effects of the transfer represent a saving of 80%. The symbols of this table have the same meaning as in Table 1.

Upon the basis of the data presented in these tables, we are able to make the following conclusions, which will be illustrated and discussed.

1. The nature of the transfer is positive for all five pairs of mazes for both humans and animals, and by all three criteria of measurement. In no instance is there any evidence of negative transfer.

The data of Tables 1 and 2 substantiate this conclusion. An

TABLE 2. AVERAGE PERCENTAGE OF SAVING IN TRANSFER.

	RATS		
	Trials	Errors	Time
A—B	77.08	85.81	83.77
A—D	69.02	79.71	90.42
A—E	19.91	54.63	63.40
A—F	63.01	42.78	59.44
A—C	57.85	46.10	34.94
	HUMANS		
A—D	51.98	94.58	88.73
A—B	67.86	88.64	67.18
A—C	19.74	20.20	29.18

examination of Table 1 shows that the averages of the original mastery for all three of the criteria are larger than for those of the transferred learning in every one of the twenty-four possible cases of comparison. It is observed in Table 2 that in every instance of measurement, in the five cases with the human subjects and the three cases with the rat subjects, a considerable percent of saving is shown to exist. It will be observed that the lowest percentage of transfer measured by trials is a fraction over 19%, by errors the lowest percentage is 20%, and by time the lowest record is 29%. From the lowest records, the percentage saved runs up as high as 77% for trials, 94.5% for errors and 90% for time.

These differences between the records for the original and the transferred learning may be due to three possible causes—chance, group differences, or the previous maze experience of the transferred group. We have evidence along two lines to prove that chance is not playing a very important role. It will be noted in Table 1 that the average deviations are rather wide, due perhaps to the small number in the groups. This fact throws some doubt on the validity of these differences. To test this we figured the Probable Difference for each of the twenty-four instances of comparison, and in every case, with but one exception, the actual difference was found to exceed the probable difference. There is the further important fact of the consistency with which the differences occur. In the eight cases of comparison of the original and transferred learning, the records for the transferred learning are uniformly the lower; this uniformity

in the records can also be observed for each of the criteria of measurement; and it obtains for the transfer stated in both relative and absolute terms. Considering the above facts, we believe that chance differences can not be regarded as the primary causal factor in the results here obtained. Further, the possibility of group differences functioning sufficiently to produce the above results is obviated by our method of selecting the groups. The rats were bought in lots of fifty to one hundred; these were mixed and the groups were selected by chance, care only being taken that both male and female subjects were in each group. Thus the possibility of having a selected group of either a good or a bad strain, a highly intelligent group, or a group of low intelligence, is eliminated. The human subjects were secured from the large number of students studying psychology in the University of Chicago, and the grouping of these was a matter of chance with the exception that both men and women were put in each group. Again, the probabilities favor our not having groups of good or bad strain, or groups of high or low intelligence. Mathematically, the chances favor the validity of the differences; the uniformity of lower records for the transferred learning enhances this probability manyfold; further, the factor of group differences was overcome by our method of securing the various groups. To our mind, the results of this experiment prove beyond doubt the existence of a positive transfer.

The positive character of the transfer is significant in view of the fact that an effort was made to so arrange the relations between some of the maze patterns as to secure a negative effect. Mazes A and B were constructed on a highly similar design with the expectation of securing a positive effect, and we were not disappointed in the results. Compare Figures 1 and 2. In designing the other four maze patterns, we aimed to secure negative results and in every case we failed to realize our purpose. The principles governing the designing of these patterns may be briefly mentioned. Maze C was so arranged that in transferring to it from A, the general direction of travel would reverse the older habits of the subjects. This fact is evident from a comparison of the patterns represented in Figures 1 and 3. Maze D

presents such an arrangement of cul de sacs, that the older habits acquired in Maze A would tend to produce many cul de sac entrances. Compare Figures 1 and 4. Simplicity, or ease of mastery governed the construction of Maze E, on the hypothesis that a transfer from a difficult to an easy maze might conduce to a high degree of confusion or disturbance. Maze F differed from A, both in conflicting arrangement of cul de sacs, and in a considerable shortening of the length of the true path. By a comparison of the patterns in Figures 1 and 6, it will be observed that in the transfer from A to F the subject must learn to omit the section numbered 6 to 10 which corresponds to a section of the true path in A. As we have stated, none of these arrangements operated to produce a negative transfer effect. However, we do not mean to assert that a negative effect between two mazes is impossible.

2. Transfer is a composite process consisting of both positive and negative elements. The acquisition of any maze may both hinder and aid in the mastery of a second maze, although the total effect is positive.

The proof of the above proposition is found in a comparison of the original and transferred learning of Maze F. In the preceding topic we noted the relation of section 6-10 in Maze F to the corresponding part of Maze A. These parts are so related that rats trained in A should possess some tendency to enter section 6-10 in the subsequent mastery of F, and naturally any undue tendency to enter these cul de sacs will be detrimental to its mastery. The hypothesis that the habits acquired in A did exert such a detrimental effect is proven by a comparison of the records of the test and control groups for this section. Rats previously mastering A—the test group—entered this section much more frequently and experienced a greater difficulty in eliminating the tendency than did the control group—animals without such training.

(1) The test group required a greater number of trials to eliminate this section. Our criterion of mastery was five successive runs without entrance. The average for the test group was 8.22, for the control group 6.15. Stating the values in rela-

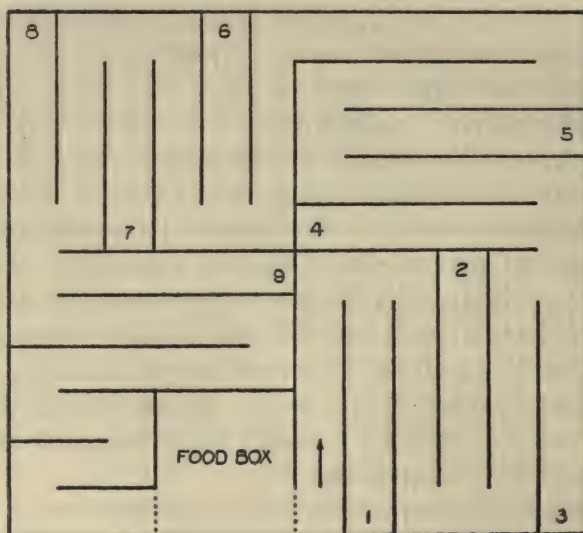


FIGURE 1: MAZE A.

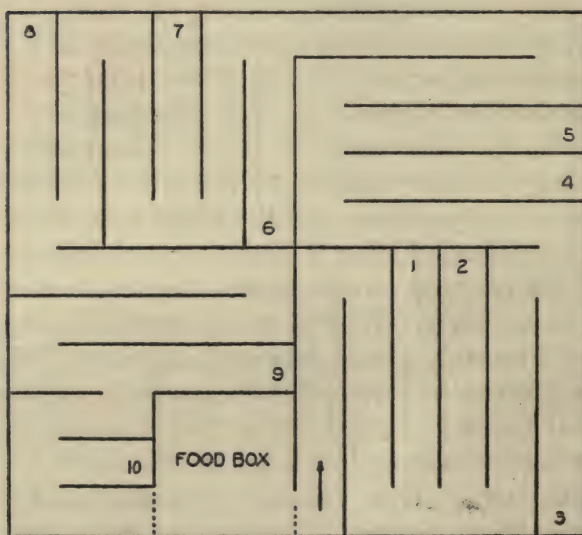


FIGURE 2: MAZE B.

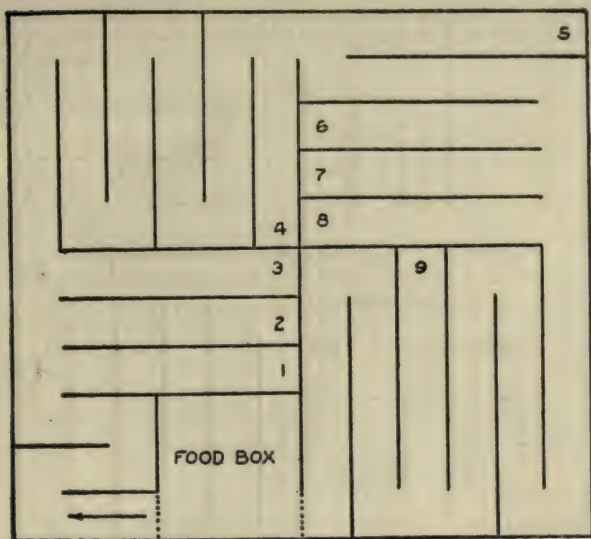


FIGURE 3: MAZE C.

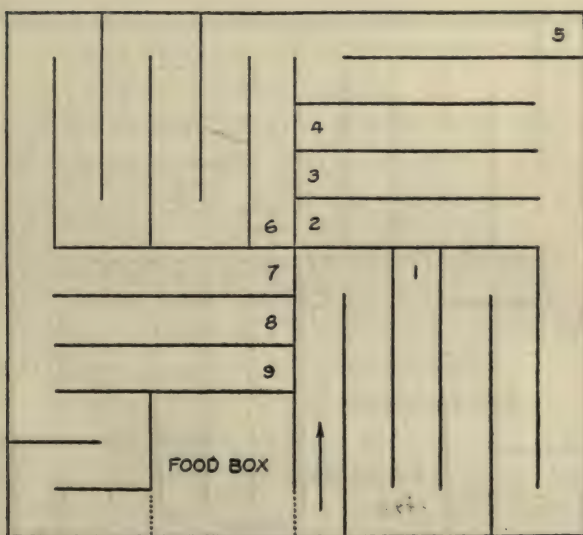


FIGURE 4: MAZE D.

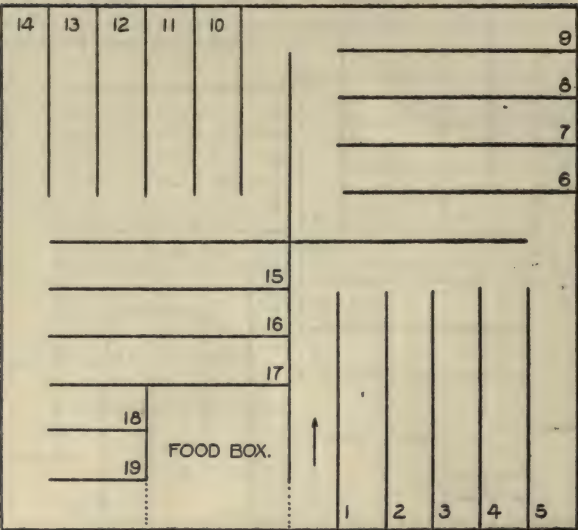


FIGURE 5: MAZE E.

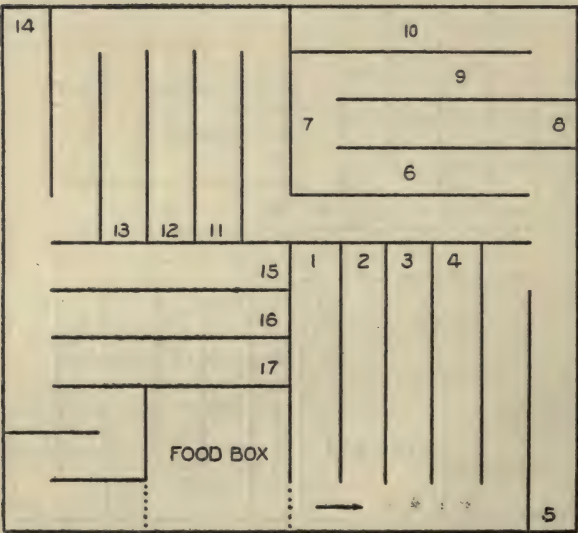


FIGURE 6: MAZE F.

tive terms, the test group required 79.5% of its total trials, while the corresponding percentage value for the control group was but 22%.

(2) The test group entered this section much more frequently. The average number of trials in which the test group entered this section was 5.88. The corresponding value for the control group was 5.15. The test group entered this section in 56% of its trials while the control group entered it but 18%. The animals frequently entered this section several times in the same trial; comparing the average number of entrances per rat, the values are 9.0 and 7.1 for the test and control groups respectively.

(3) The test group made the greater number of errors in this section. Since our unit of error is a single runway, it was possible for a rat to make a number of errors in a single entrance. The average number of errors per rat for the test and control groups were 30.4 and 24.0. Of the total number of errors made by the test group in mastering the maze 41.2% were due to this section. The corresponding value for the control group was 10.8%.

3. The third feature to be noticed from the above data is the fact that the degree of transfer is in part a function of the activities set up in the second maze situation.

Maze A is the constant activity, and the other five mazes are the varying processes of the second problem. If the transfer effect is mainly a function of the constant activity, the amount saved should be fairly uniform; the differences should be only such as can be accounted for by mere chance, or slight individual differences. On the other hand, a wide divergence of results will indicate that the degree of transfer is in part a function of differences in the character of the second maze.

Upon inspection of Table I, it is found that the amount saved is not uniform, but rather that the range of the variations is quite wide. We discover wide variations in the average amount saved measured by trials, errors and time. The average saving for trials in the rat records vary from .9 to 43.3, and in the human records from 6 to 22.8. The averages for errors with

the rats vary from 10.4 to 192.2, and with the humans from 26.9 to 252.8. Similar results are observed in the time records; the rats vary from 135.4 to 2511.1, and the humans from 214.9 to 1016.6. Likewise, the results, stated in relative terms (see Table 2), indicate that the percentage of transfer varies with the change of the relation between the two activities.

Two possible explanations can be offered to account for these marked variations in the degree of transfer; either these are due to mere chance, or they are a function of the second activity.

The dependence of the degree of transfer upon the second maze activity is proven by the fact that the actual amount saved in each case is roughly proportionate to the original learning records of the five mazes. This fact is evident from an inspection of the data of Table 1. Considering the results for trials in the rat records, the largest saving (43.3 trials) was secured in the transfer from A to B, and B was the most difficult of the five mazes in terms of number of trials necessary to master. Likewise, the smallest amount saved (.9 trials) was secured in transferring to E, and this maze presented the least difficulty in original mastery. This relation between amount saved and the original learning records for any maze is also evident from Table 2, which states the actual amount saved for any maze in percentage terms in reference to the data for original learning. If the amount saved depended absolutely upon the difficulty of the second maze, these percentage data should be exactly the same for the five pairs of mazes. Provided there is only some degree of correlation between the two sets of data, uniformity of percentage data would not obtain, but the range of divergence for these relative data should be much less than that for the data of Table 1, giving the amount saved in absolute terms. Stating the transfer effect in relation to the corresponding learning data (Table 2), does decrease the range of divergence. For example, the saving for the rats varies between the extremes of .9 and 43.3 trials, and this divergence is decreased to 19.91 and 77.08 when estimated in relative terms; a better example is found by comparing the transfer to B with that to F. The absolute records represent a saving of 43.3 and 17.6 trials respectively, while the corresponding percentage records are 77.08 and 63.01.

The dependence of the amount saved upon the original learning records of the second maze, is also proven by the fact that a positive correlation obtains between the two sets of data. The discussion of this correlation phenomenon will be reserved for the succeeding topic.

4. A positive correlation exists between the degree of transfer and the difficulty of the second problem.

In the preceding topic, it was maintained that the variations in the amounts saved is in part a function of the second activity. In supporting this statement, we offered as evidence the fact that the differences in saving are correlated with the relative difficulty of the second problem. We have computed the correlation between the degree of transfer and the difficulty of the second maze, between the amount of effort saved due to transfer and the amount of effort expended in the original learning. The materials for this correlation are easily obtained from Table 1. The correlations were computed by the ranking method, and Table 3 presents the results for trials, errors, and time.

TABLE 3. CORRELATION BETWEEN DEGREE OF TRANSFER AND DIFFICULTY OF SECOND MAZE.

	Trials	Errors	Time
Rats	1.00	.70	.70
Humans	.50	1.00	1.00

A positive correlation obtains in the table by all three of the criteria and for both humans and animals. With the rats a perfect correlation exists for trials, and plus .70 for errors and time. In the human records we find a perfect correlation in the matter of errors and time, and a value of plus .50 by the criterion of trials.

In view of the fact that these values are based upon but five mazes for the animals, and upon three for the humans, one can not credit the validity of any single value, especially those as low as .50. The validity of the results must depend rather upon their *uniformity*. A positive correlation was obtained in every one of the six cases, and this fact enhances the probability of the validity of any single measurement manifold. On the other hand, one must not be imbued with scepticism because a

perfect correlation was not obtained in every case. Our conclusion merely states that the degree of transfer depends *in part* upon the difficulty of the second maze. Subsequent records will prove that the transfer effect is also a function of the first maze, and hence a perfect correlation with the data for either of the two mazes is not to be expected.

5. A positive correlation exists between the degree of transfer and the similarity of two maze patterns.

The main difficulty in a comparison of this kind concerns the measurement of the degree of similarity between any pair of mazes. One of the current theoretical explanations of the phenomenon of transfer is stated in terms of the partial identity of the neural elements existing between the two activities. Any measurement of such a relation between two activities is necessarily impossible.

In this experiment, we employed two methods of measuring the similarity of a pair of maze patterns. The first method used was that of 'order of merit' or 'relative position.' Nineteen individuals, who understood the maze problem, were asked to rank the five mazes in order as to their similarity to Maze A. They were asked to judge the similarity upon the basis of the two factors of the positional relation of the true pathway and the cul de sacs, and the direction of the course of travel. The results of these nineteen judgments were as follows: B was placed in first place 17 times and in second place twice; D was ranked second 17 times and first two times; C was given fifth rank by 14 and third rank by 5. The most difficult task concerned the ranking of E and F. The majority (11 out of 18), however, gave F the preference. From these judgments the mazes were ranked as follows: B-1, D-2, F-3, E-4, and C-5. The rankings as to the degree of transfer were determined from Table 1.

The results of the computations are given in Table 4. It will be noted that the correlation is positive by each of the criteria for both humans and rats. As regards trials there is the result of plus .30 for the rats and plus .50 for the humans. The record by errors gives a correlation of plus .70 for rats and 1.00 for

TABLE 4. CORRELATION BETWEEN AMOUNT SAVED IN TRANSFER AND SIMILARITY OF MAZES.

	Trials	Errors	Time
	Similarity by Order of Merit.		
Rats	.30	.70	.60
Humans	.50	1.00	1.00
	Similarity in Terms of Difficulty of Mastery.		
Rats	.70	.90	.30
Humans	.50	.50	1.00

humans; in the time records the rats have a correlation of .60 and the humans 1.00. Some of the values are high enough to indicate a valid correlation. Other values are so low that little importance can be attached to their significance when regarded singly. The validity of the results must depend primarily, not upon individual instances, but upon their uniformity. Some degree of positive correlation obtains for all six measurements, and this fact enhances their probable significance manifold.

The second method measured the similarity of a pair of mazes in terms of their relative difficulty of mastery. If Maze A required forty trials for its mastery, while Mazes B and C were mastered in thirty and fifteen trials respectively, it is evident that A and B are more similar in respect to difficulty than are A and C. The five pairs of mazes can thus be ranked as to relative difficulty in terms of trials, errors, and time from the comparative data of Table 1.

These correlations are also given in Table 4, and a positive value again obtains for the six comparisons. The validity of these correlations again depends upon their uniformity.

We also compared the ranking as to similarity by the first method with the three sets of ranks obtained by the second method. The correlation value for trials was minus .30, while positive values of .50 and .60 were obtained for errors and time respectively. This indicates that any judgments as to the relative similarity of pairs of maze problems will anticipate more correctly their relative difficulty when measured by time and errors than by the factor of trials.

6. The transfer results are somewhat similar for the human and animal groups. Evidently the laws and conditions governing transfer are not radically different in the two organisms.

Transfer obtained for both human and animal subjects and its character was positive or beneficial in both cases. However, the rats evidenced more ability in utilizing a previous maze experience in a new situation, which is proven by the fact that the rats effected a larger saving in the transfer. The difference is quite pronounced and obtains for each of the three pairs of mazes when the results are stated in either absolute or relative terms. This is evident from an examination of the comparative data of Table 5. Rats and humans differ little, however, when the transfer is measured by the saving of errors, but the difference such as it is favors the human subjects. The humans effected the greater saving of errors for two of the three pairs of mazes for both the absolute and relative methods of stating the results. Thus, on the whole, the rats were able to profit more from their previous tuition in Maze A when the effect is measured in terms of trials or time. The rats expended the greater amount of trials and time to master A and made the greater saving of trials and time because of this previous training in A. The humans, on the other hand, required the greater number of errors to master A, and likewise profited the most from this tuition when its effects are measured in terms of errors. This fact will be observed by a comparison of the records for the original learning of Maze A as given in Table 5. This difference between humans and animals as to the amount saved due to their previous experience in A would seem to be a result of the character of that training. In other words, the amount of

TABLE 5. COMPARATIVE RESULTS OF HUMANS AND RATS
Absolute Amounts Saved.

	Trials		Errors		Time	
	Rats	Humans	Rats	Humans	Rats	Humans
A—B	43.3	22.8	192.2	252.8	2068.1	1016.6
A—C	27.4	7.5	109.2	26.9	1037.2	214.6
A—D	11.5	6.	122.4	192.	2511.1	686.3
	Percentage		Amounts Saved.			
	Rats	Humans	Rats	Humans	Rats	Humans
A—B	77.08	67.86	85.81	88.64	83.77	67.18
A—C	57.85	19.74	46.10	20.20	34.94	28.19
A—D	69.02	51.98	79.71	94.58	90.42	88.73
Records for Original Learning of Maze A.						
	38.9	24.3	205.9	231.2	1782.4	970.8

transfer would seem to be in part a function of the character of the training secured in the previous maze, a proposition which will be further demonstrated by the succeeding experiment. While humans and animals manifest some difference as to the amount saved, yet they are similar in the following respects: that pair of mazes which induces the greatest amount of saving for rats has a like effect with humans; mazes which give the least effect with humans produce similar results with the rats. This formulation is true for the transfer stated in either absolute or relative terms; the truth of the proposition may be determined by computing the correlation between the amounts saved for the two classes of subjects. These correlation values were determined separately for trials, errors, and time when stated in both absolute and relative terms. The rankings were secured from Tables 1 and 2 and the results are to be found in Table 6.

TABLE 6. CORRELATION OF TRANSFER BETWEEN HUMANS AND ANIMALS.

	Trials	Errors	Time
Absolute	1.00	1.00	.50
Relative	1.00	.50	1.00

It is observed from the above table that there is a perfect correlation for trials in both absolute and relative terms; for errors we have a perfect correlation in absolute terms and plus .50 in relative terms; for time the results are 1.00 in relative terms and plus .50 in absolute terms. Thus we observe that in six cases of comparison there are four perfect correlations, and two of only fifty. However, we again put forth the statement that we must not judge by a single case, but by all the cases. So the consistency of the positive results strengthens the probability of a high correlation between human and animal learning as regards the phenomenon of transfer.

These facts prove not only that transfer obtains for both the human and animal realms, but that human and animal organization is highly similar so far as the laws and conditions of transfer are concerned. It further indicates that the processes involved are highly similar, and that no factors, such as rational activities peculiar to the human subjects, are functioning in this process of transfer.

7. A positive correlation exists between any two of the three criteria of measurement.

The efficiency of the criteria used in measuring any part of the learning process, is a question of great importance. In measuring the transfer effect in this study we have employed three criteria, trials, errors, and time. The interdependent relation existing between any two of these criteria can best be determined by correlation. The correlation values were computed from the records of the individuals in each of the transferred groups. The ranking method was used, and the results are given in Table 7.

TABLE 7. CORRELATION BETWEEN CRITERIA OF MEASUREMENT.

	RATS		
	Trials-Errors	Trials-Time	Errors-Time
A-B	.48	.03	.92
A-C	.34	.15	.87
A-D	.76	.76	.94
A-E	.83	.88	.76
A-F	.67	.74	.93
	HUMANS		
A-B	-.10	.40	.70
A-C	.93	.56	.37
A-D	.78	.83	.88

From this table of data, the following comparisons can be made. (1) The correlation is positive in twenty-three of the twenty-four cases. This uniformity indicates the existence of some dependent relation between trials and errors, trials and time, and errors and time as a means of measuring the transfer. (2) The human subjects exhibit the higher values in five of the nine cases of comparison. Whether this is due to chance or represents a general tendency, it is impossible to say at this time. (3) A higher correlation obtains between time and errors than between trials and either of the other two criteria in six of the eight possibilities. But apparently no difference obtains for trials and errors as compared with trials and time. (4) The pair of mazes A-D gives the highest values in five of the six cases. No uniform difference obtains between A-B and A-C.

B. DEPENDENCE OF TRANSFER UPON THE CHARACTER OF THE FIRST PROBLEM.

The object of the second experiment is to determine the dependence of the degree of transfer upon the nature of the first maze.

The possibility of the degree of transfer depending upon either the first or second learned problem has been previously mentioned. This was illustrated from the discussion about the Classics. The student may learn mathematics, history and science first and then study Latin; the degree of transfer in this instance might depend, not upon Latin, but upon the previously mastered subject. This is the type of problem to be discussed in this section of our paper.

In order to test experimentally such a proposition as was raised in the preceding paragraph, one factor must be kept constant. Maze A is again the constant activity. One group of subjects first learned Maze B and then was transferred to Maze A. Another group was transferred from C to A, one from D to A, a fourth from E to A, and a fifth from F to A. The varying factor in this situation is the first learned maze; the second acquired problem is kept constant for all of the groups. If the degree of transfer is wholly a function of the second or constant activity, the variations should be only such as can be accounted for in terms of group, or mere chance differences. Should marked variations obtain in the degree of transfer, it can be said that these are due in part to the differences in the character of the first learned problem.

The amount of transfer was measured in the manner described in the preceding topic, viz., by the comparison of the transferred learning with the original mastery of the same maze. A control group for Maze A is therefore necessary, and this group contained 54 rats and 21 humans. The computations were figured in the same manner as described above for the three criteria of trials, errors, and time.

Table 8 presents the results in the absolute terms of group averages. These group averages, together with the average deviations, for the original and the transferred learning are

TABLE 8. COMPARATIVE RECORDS OF ORIGINAL AND TRANSFERRED LEARNING.

RATS			
	Trials	Errors	Time
<i>A</i>	38.9±12.9	205.9± 60.9	1782.4±821.8
<i>B-A</i>	7.7± 6.6	9.7± 9.5	221.3±212.5
Sav.	31.2	196.2	1561.1
<i>C-A</i>	19.3± 8.9	76.3± 28.6	487.1±218.6
Sav.	19.6	129.6	1295.3
<i>D-A</i>	23.7±11.7	58.8± 39.2	457. ±199.7
Sav.	15.2	147.1	1325.4
<i>E-A</i>	22.5±11.6	82.3± 36.8	651.5±318.
Sav.	16.4	123.6	1130.9
<i>F-A</i>	35.1±10.9	99.2± 24.9	704.7±254.4
Sav.	3.8	106.7	1077.7
HUMANS			
<i>A</i>	24.3± 9.7	231.2±111.6	970.8±388.2
<i>B-A</i>	5. ± 4.8	5.8± 5.7	118.6±104.7
Sav.	19.3	225.4	852.2
<i>C-A</i>	25.5± 8.1	171.7±106.3	670.2±380.8
Sav.	-1.2	59.5	300.6
<i>D-A</i>	27.5±10.	128.5± 57.1	524.3±153.8
Sav.	-3.2	102.7	446.5

given for each of the criteria. The letters in the table indicate the records for the several mazes; thus *A* denotes the record for the original learning of that maze, and *B-A* indicates the transferred learning of the same maze. The *B*, *C*, *D*, *E*, *F* in connection with *A*, e. g. *C-A* etc., means that the previous maze experience of the group consisted of the maze shown by that letter. The symbol Sav. refers to the average amount saved for each group and these figures thus measure the transfer effect for each group.

From an examination of the data in this table, we are able to make the following conclusions, which will be illustrated and discussed in order.

I. The nature of the transfer is positive for all five mazes. The above statement is true for twenty-two of the twenty-four

instances of comparison. The two instances that failed to show positive transfer are found in the human records for trials in the transfer from *C-A* and *D-A*, and yet in these two cases a positive transfer obtains in terms of errors and time. In the transfer from *C* to *A* a loss of practically 5% is noted by the criterion of trials, and in the transfer from *D* to *A*, by the same criterion, a loss of over 13% is observed. The cause of this is not known; it may represent a matter of chance, or it may be a valid instance of negative transfer. In the transfer from *C* to *A* the loss is rather small, and a saving of over 40% is shown by the other two criteria, and this fact might argue that chance is responsible. One consideration may be offered, which tends to indicate that this loss is due to an individual peculiarity. In this transferred group, *D-A*, one subject used more trials to learn Maze *A* than did the other twenty-one subjects. This same person was also the most erratic subject in the original learning of Maze *D*; he used next to the largest number of trials, made considerably more errors, and consumed much more time than did any of the other subjects. The average deviation for his group in the transferred record *D-A* is larger than in the original learning of *A*, and this fact, we believe, can be attributed to this individual's record. Judging from these considerations, it appears to us that the results prove, with a high probability, that positive transfer exists in the cases *C-A* and *D-A* with human subjects.

The three possible causes for the differences between the records of the original and the transferred learning, that were suggested in discussing topic *A*, are equally relevant here. The causal factors may be chance, group differences, or the previous training of the transferred groups. The arguments previously advanced to prove that the differences are real and represent a transfer effect have an equal bearing in this connection. It is noted that the average deviations are rather wide, and this result is due perhaps to the small number of subjects in the various groups. For the data of this table we computed the Probable Difference, and in twenty-one of the twenty-four instances of measurement the actual difference is found to ex-

ceed the probable difference. We have the same situation as regards the matter of consistency; the differences are uniformly lower for the transferred records in twenty-one of the twenty-four instances of measurement. This consistency strengthens the probability, to a high degree, that chance is not responsible for the differences. The groups used in this experiment were secured and mixed as described in the previous experiment, thus eliminating the possibility of the differences being due to group variations. Further the control group is composed of a rather large number of subjects. Every subject that originally learned Maze A during the eight months of the experimentation was utilized in this control group. This fact lessens to a greater degree the possibility that group differences are functioning in any large part. We have thus shown by mathematical calculation and by the matter of consistency, that chance can not be functioning to a significant degree; the possibility of group differences being a large causal factor is eliminated by our method of securing and mixing the subjects of the various groups. Hence, we again believe, beyond doubt, that the nature of the transfer is positive.

In this, as in the former experiment, the transfer effect remained positive, despite our efforts to so arrange the maze patterns in such a way as to produce a negative effect. These relations between the maze patterns have been described and illustrated. The transfer from C to A reverses the direction of travel; D and A involve a new arrangement of cul de sacs. E and A represent a transfer from a simple to a difficult pattern, while F to A involves an increase in the length of the true pathway and a new arrangement of cul de sacs.

2. The degree of transfer in this experiment is in part a function of the first problem.

As we have previously stated, the second maze is the constant activity, while the first maze is different for each pair. If the degree of transfer is a function of the constant factor, the differences should not vary beyond what can reasonably be attributed to mere chance. On the other hand, if the degree of transfer varies beyond the possibility of mere chance differences, this re-

sult must be regarded as a function of the first or varying activity.

By an inspection of Table 8, it will be observed that the amount saved in absolute terms varies considerably in the five cases with the rats and in the three cases with the human subjects. The records of the rats vary by trials from 3.8 to 31.2, and those of the humans from -3.2 to 19.3. For errors rats have a range in the saving varying from 106.7 to 196.2, and the humans from 59.5 to 225.4. In the time records there exist equally wide variations; those of the rats vary from 1077.7 to 1561.1, and those of the humans from 300.6 to 852.2.

The previous discussion (page 24) of the significance of such wide variations in the degree of transfer is in point here. We believe that the possibilities are rather remote that mere chance would cause such a wide range of variations as noted. By comparing the amounts saved (Table 7) with the figures representing the original mastery of the five mazes (Table 1), it will be noted that the two sets of data are roughly proportionate. For example, the largest saving in trials for the rats was obtained in the transfer from B and this maze required the greatest number of trials for its mastery. Likewise C mediated the next highest amount of saving in trials and this maze was also second as to difficulty of mastery. The dependence of the amount saved in each case upon the difficulty of the first maze is best demonstrated by computing the correlation between the two sets of data. By referring to the correlation values in Table 9 below, it is observed that a positive correlation exists for trials, errors, and time, and for both humans and rats. This fact proves that as changes in the character of the first maze were made, corresponding changes in the degree of transfer occurred. Hence, we believe that we are justified in concluding that the varying degree in the transfer effect is in part a function of the varying activity, viz. the first learned problem.

3. A positive correlation is found to exist between the degree of transfer and the difficulty of the first problem.

The values of Table 9 represent the correlation between the amounts saved in the transfer to Maze A and the amounts of

effort, figured by trials, errors, and time, expended in the original mastery of Mazes B, C, D, E, and F. The ranking method of computation was employed.

TABLE 9. CORRELATION BETWEEN DEGREE OF TRANSFER AND DIFFICULTY OF FIRST PROBLEM.

	Trials	Errors	Time
Rats	.60	.60	.50
Humans	.90	1.00	1.00

On examining Table 9, it is noted that the correlations are positive in every instance for both humans and rats. The data between which the correlation was computed in each case were so few in number that but little reliance can be placed upon the validity of any one of the correlation values. In estimating the validity of these values, the uniformity of positive results must be emphasized. A positive correlation was secured in each of the six comparisons, and this consistency increases the probability of any single value manyfold. The validity of the values can not be attacked because a perfect correlation was not obtained in each case, for such is not to be expected as we demonstrated in our first experiment, that the amount of transfer is also correlated with the difficulty of the second of each pair of mazes.

The correlation values are noticeably higher for the human subjects than for the rats. It is hardly possible to determine whether this difference is due to chance, or to some difference between human and animal organisms.

4. A positive correlation is found between the amount of transfer and the degree of similarity of the two maze patterns.

TABLE 10. CORRELATION BETWEEN AMOUNT SAVED IN TRANSFER AND SIMILARITY OF MAZES.

	Trials	Errors	Time
Similarity by Order of Merit.			
Rats	.10	.60	.60
Humans	.50	1.00	1.00
Similarity in Terms of Difficulty of Mastery.			
Rats	.10	.80	.00
Humans	.50	.50	1.00

The degree of similarity between any two maze patterns was

measured by the same two methods as were used in the first experiment. In fact the rankings as to similarity are the same in both experiments. The correlation values are given in Table 10.

A positive value was secured in 11 of the 12 comparisons, and these positive values range from .10 to 1.00. The validity of our conclusion must again depend upon the uniformity of the results. The values are higher for the humans in five of the six cases, and in general errors and time give higher values than does the criterion of trials. Since the rankings as to similarity are the same as in the first experiment, we may repeat the former statement that some degree of positive correlation obtains between the two methods of estimating the similarity.

5. A comparison of human and rat results reveals an essential similarity. No radical difference can be inferred from these data.

Both classes of subjects exhibit a positive transfer. Two exceptions are to be found in the human records, but these are probably due to chance or individual peculiarity. In this experiment the rats exhibited the greater amount of transfer in eight of the nine instances of comparison. The single exception refers to the saving in errors due to the transfer from Maze B to A. When the amount saved is stated in terms of its percentage relation to the records representing the mastery of the first maze, the rats again manifest the greater effect from their previous tuition in seven of the nine comparisons. While the rats were able to utilize their previous training to a greater degree than the humans, yet that maze which gave the highest transfer for the rats also yielded the highest value for the humans. On the other hand the lowest values for both humans and rats were secured from the same maze. A perfect positive correlation obtained for each of the three criteria, and when the transfer is stated in both absolute and relative terms.

6. The results indicate a positive correlation between any two of the three criteria of measurement.

As in the previous instance of determining such a correlation, we have computed the values from the records of the individuals

in each of the transferred groups. The ranking method of correlation was employed, and Table II presents the results upon which the above conclusion is made.

Comparisons, such as were made in the first experiment, can also be made from the data in this table. (1) There is a positive value in all of the twenty-four cases, and this again indicates the existence of some dependent relation between any two of the three criteria as a means of measuring the transfer. (2) The human subjects exhibit higher values in eight of the nine comparisons. We are unable at this point to determine whether this is due to chance, or whether a general tendency is present. (3) Little uniformity obtains as to the matter of higher values between errors and time and either of the other two criteria. Likewise, there is a lack of uniformity in comparing the results for trials and errors, and trials and time. (4) No definite statement can be made as to uniformly higher values for any one pair of mazes.

TABLE II. CORRELATION BETWEEN CRITERIA OF MEASUREMENT.
RATS

	Trials-Errors	Trials-Time	Errors-Time
B—A	.78	.98	.80
C—A	.61	.67	.75
D—A	.97	.97	.93
E—A	.95	.55	.62
F—A	.48	.88	.29
HUMANS			
B—A	1.00	1.00	1.00
C—A	.83	.83	1.00
D—A	.99	.90	.95

C. DEPENDENCE OF AMOUNT SAVED UPON DIRECTION OF TRANSFER.

The purpose of this section is to determine whether or not the amount saved is in part a function of the direction of transfer between any pair of mazes.

The problem is readily stated in terms of mazes. In one case the transfer is from A to B, while a second group is transferred from B to A. The same pair of mazes is employed in both tests; the experiments differ only in what we have termed the direction

of the transfer. If the amount saved is different for the two cases, the direction of the transfer must account for the result. If the direction of transfer is not a determining factor, the degree of saving should be practically the same for both experiments.

No new experimental data are required for the solution of this problem, as the first experiment furnishes the results from groups that were transferred from Maze A to each of the other five mazes, while the second experiment gives us the data for the opposite direction of transfer for the same pairs of maze problems. The results from the two previous experiments, found in Tables 1, 2, 7, and 8, are thus utilized and the data so arranged in Tables 12 and 13 as to facilitate a comparison of the amounts saved for the two directions of transfer. Separate comparisons are made for the human and rat subjects, for each of the five pairs of mazes, and for each of the three criteria of measurement. Table 12 states the amounts saved due to transfer in absolute terms, while in Table 13 a comparison is made of the saving stated in relative or percentage terms. The symbol A-B means that the subjects were transferred from A to B, while B-A refers to the opposite direction of transfer for the same pair of mazes. The symbol Dif. indicates the difference in the amounts saved in the two directions of transfer.

1. The direction of transfer is in part a deciding factor in determining the degree of transfer.

The differences between the two sets of results may be due to group peculiarities, chance, or the direction of transfer. The possible functioning of group peculiarities has been obviated by our method of group selection. While differences exist for all twenty-four cases of comparison, yet such a result would naturally be expected even though chance were the only factor operating. Neither are the differences in the majority of the cases large enough to exclude the possibility of chance. The influence of the direction of transfer in mediating the differential results is proven by the fact that the differences are a function of the degree of similarity of the various pairs of mazes. That pair of mazes possessing the highest degree of

similarity yields the smallest difference of saving when the direction of transfer is reversed, while the largest difference of saving tends to obtain for the most dissimilar pair of mazes. The size of the differences for the various pairs of mazes is thus not entirely a matter of chance; it depends to a slight extent upon the degree of similarity of the maze patterns. To prove this relationship, we have computed the correlation between the size of the differences and the degrees of similarity of the various pairs of mazes. From the data of Table 12, we ranked the pairs in the order of increasing values of differential results, and correlated this order with those representing their degree of

TABLE 12. COMPARATIVE AMOUNT SAVED FOR TWO DIRECTIONS OF TRANSFER.
RATS

	Trials	Errors	Time
A—B	43.3	192.2	2068.1
B—A	31.2	196.2	1561.1
Dif.	12.1	4.0	507.0
A—C	27.4	109.2	1037.2
C—A	19.6	129.6	1295.3
Dif.	7.8	20.4	258.1
A—D	11.5	122.4	2511.1
D—A	15.2	147.1	1325.4
Dif.	3.7	24.7	1185.7
A—E	.9	10.4	135.4
E—A	16.4	123.6	1130.9
Dif.	15.5	113.2	995.5
A—F	17.6	54.7	719.1
F—A	3.8	106.7	1077.7
Dif.	13.8	52.0	358.6

HUMANS

A—B	22.8	252.8	1016.6
B—A	19.3	225.4	852.2
Dif.	3.5	27.4	164.4
A—C	7.5	26.9	214.9
C—A	—1.2	59.5	300.6
Dif.	8.7	32.6	85.7
A—D	6.0	192.0	686.3
D—A	—3.2	102.7	446.3
Dif.	9.2	89.3	239.8

similarity as determined by the two methods described in the first experiment—similarity of difficulty, and similarity of maze patterns. For the rat records, positive values of .30, 1.00, and .20 were secured for trials, errors, and time, respectively when the degree of similarity was measured in terms of difficulty of mastery.

TABLE 13. COMPARATIVE PERCENTAGES SAVED FOR TWO DIRECTIONS OF TRANSFER.

	RATS		
	Trials	Errors	Time
A—B	77.08	85.81	83.77
B—A	80.11	95.27	87.58
Dif.	3.03	9.46	3.81
A—C	57.85	46.10	34.94
C—A	50.38	62.94	72.67
Dif.	7.47	16.84	37.73
A—D	69.02	79.71	90.42
D—A	39.14	71.43	74.36
Dif.	29.88	8.28	16.06
A—E	19.91	54.63	63.40
E—A	42.12	60.03	63.44
Dif.	22.21	5.40	.04
A—F	63.01	42.78	59.44
F—A	9.59	51.81	60.46
Dif.	53.42	9.03	1.02
	HUMANS		
A—B	67.86	88.64	87.18
B—A	79.41	97.49	87.78
Dif.	11.35	8.85	.60
A—C	19.74	20.20	29.18
C—A	—4.98	25.74	30.97
Dif.	24.72	5.54	1.79
A—D	51.98	94.58	88.73
D—A	—13.21	44.42	45.99
Dif.	65.19	50.16	42.74

The corresponding values for the human subjects were .50, —.50, and —.50. The mazes were also ranked as to differential results by taking an average of the values for the three criteria, and correlation values of .67 and .50 were secured for the rat and human subjects respectively. The differential results were also

correlated with the degree of similarity determined by the method of the 'order of merit,' and the following values were obtained: rats, .20, .40, and —.10 for trials, errors, and time respectively; the corresponding values for humans were .50, .50, and —.50; when the order of differential results was determined by averaging the values for the three criteria, values of .10 and .50 were secured for the rats and humans respectively. The mazes were again ranked in order of increasing values of the differential results as given in percentage terms in Table 13. This system of values was likewise correlated with the degree of similarity of the maze patterns. These correlation values were practically identical with those above and so need not be given. It is noted that small positive values predominate; positive values were secured in 14 of the 17 computations. The validity of our proposition must depend upon the consistency with which these positive values were secured. To our mind these data prove that the direction of the transfer between any pair of mazes exerts some slight effect upon the resulting degree of saving.

2. The relative amount of retracing differs according to the direction of transfer.

The data supporting this conclusion are found in Table 14. We recorded separately the errors due to entrances into cul de sacs, and returns over the true path. The table gives in percentage terms the number of retracing errors relative to the total number of errors made in the mastering of each maze. For example, in the mastery of B by the rats, 41.7% of the total number of errors was due to retracing. The corresponding value for the transferred learning of B is 60.7%. In this case the transfer increased the relative amount of retracing, and this fact is denoted in the table by the positive sign plus placed after the value 60.7%. A different result was obtained for the opposite direction of transfer. The percentage values are 40.3 and 27.8 respectively for the original and the transferred mastery of A. Transfer in the direction of B-A thus decreased the relative number of errors due to retracing, and this fact is indicated in the table by the minus sign placed after the value 27.8. Direction of transfer thus operates differently for the mazes A

and B; one direction increases the relative amount of retracing while the other increases it. An inspection of the positive and minus signs for the various pairs of mazes reveals the fact that this differential effect due to the direction of transfer obtains for four of the five comparisons of rat records and for two of the three cases for the human subjects.

TABLE 14. PERCENTAGE OF ERRORS DUE TO RETRACING.

Rats		Humans	
B 41.7	A 40.3	B 65.8	A 69.3
A-B 60.7+	B-A 27.8—	A-B 59.8—	B-A 27.6—
C 66.7	A 40.3	C 48.4	A 69.3
A-C 76.3+	C-A 38.3—	A-C 44.2—	C-A 69.6+
D 71.5	A 40.3	D 60.3	A 69.3
A-D 75.3+	D-A 17.2—	A-D 60.9+	D-A 64.8—
E 54.7	A 40.3		
A-E 41.8—	E-A 36.1—		
F 33.7	A 40.3		
A-F 35.1+	F-A 30.1—		

3. The influence of the direction of the transfer is essentially identical for human and animal subjects.

The correlation values given above (page 43 f.) are essentially similar for the two types of subjects. It will be noted in Table 12 that the direction B-A gives for both humans and rats a smaller saving in trials than does the reverse direction of A-B. In six of the nine possible comparisons of this sort, those directions giving the greater saving for the rats also give the larger values for the human subjects. This correlation is perfect when the percentage values of Table 13 are utilized in the comparison. Evidently the conditions governing transfer are essentially the same for the two types of organism.

D. LOCUS OF THE TRANSFER IN THE LEARNING CURVE.

Our purpose here is to compare the original and transferred learning curves, thus enabling us to discover what part of the learning curve is affected by the transfer.

Transfer reduces the total amount of time and the total number of errors in learning a maze. This reduction may be accom-

plished in one of three ways. (1) The saving may be distributed proportionately among the various trials. In this case the curves for the transferred and the original learning would be similar in form but different only in height. For example, the number of errors made in each trial in transfer may be one half of that made in the original mastery of the same maze. (2) The two curves may be identical in height and form for the first trials and differ only in the final trials. In this case the saving due to transfer would be confined to the final stages of mastery, and the curve for the transferred learning would exhibit a sudden final drop. (3) The saving due to transfer may also be confined to the early trials. In this instance, the curve for the transfer would begin with much lower values than the curve for the original mastery, decrease much more gradually at first, and finally become identical with that for original learning in the last stages of mastery. The initial drop characteristic of the normal maze curve would thus be absent in the transfer curve. The term 'locus of transfer' will be used to represent that group of trials in which the transfer effect is mainly manifested.

The locus of transfer is on the average confined to the first five trials. Subjects transferred to any maze are saved the equivalent of the first five trials of effort; they begin the problem at an advanced stage of mastery and complete it in a normal manner. The transferred curves thus do not exhibit that sharp initial drop characteristic of normal curves.

This general conclusion is illustrated by typical data found in Figures 7 and 8. To the left of each figure are found the error and time curves representing the progress of learning for the first five trials of the original mastery of the various mazes. To the right is placed the initial part of the corresponding curves for transferred learning. It will be noted that the transferred curves begin at a level closely approximating that reached by the curves for the original learning at the fifth trial. These particular curves have been selected for purposes of illustration, because they represent not only the average but the most frequent in the sixteen cases of comparison. Twelve of the sixteen cases closely approximate the conditions represented by these figures.

50

40

30

20

10

1

2

3

4

5

HUMANS

B

A-B

1

2

3

4

5

50

40

30

20

10

1

2

3

4

5

RATS

B

A-B

1

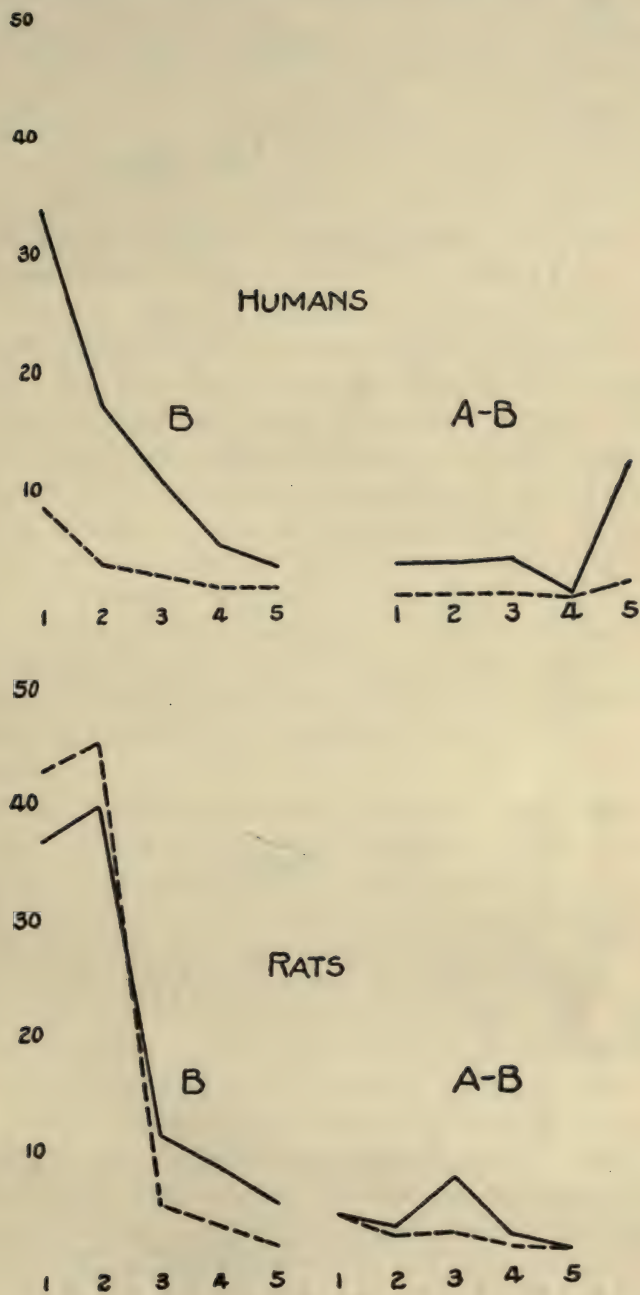
2

3

4

5

FIGURE 7.



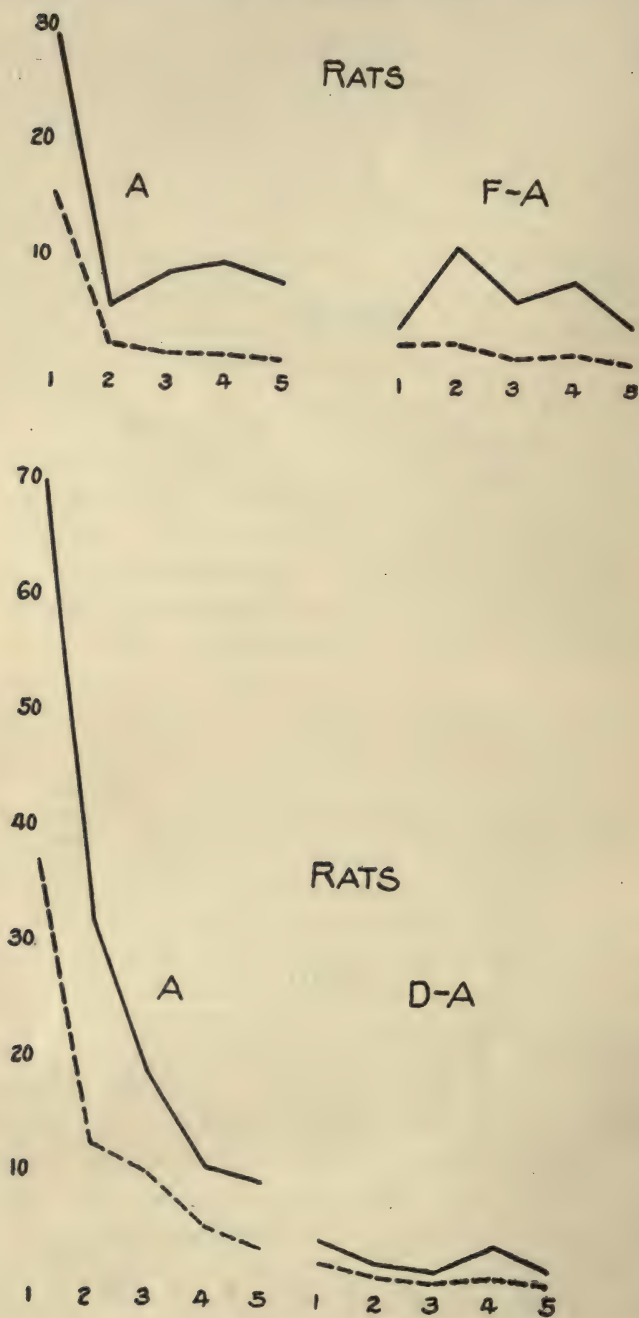


FIGURE 8.

Four cases diverge from the type. In C-A for the humans and A-F for the rats, the two curves are more nearly identical in form. B-A and A-D for the humans represent the opposite divergence from the type; here the saving due to transfer is equivalent to seven or eight trials.

E. SELECTIVE EFFECT OF TRANSFER UPON TYPES OF ERROR.

Transfer effects a saving in the total number of errors necessary to master a maze. These errors comprise two sorts—those due to entering cul de sacs, and those due to retracing in the true pathway. We have listed these two types of error separately, and have computed the degree of saving for each. This was done by dividing the number of errors made in the original mastery of the maze into the number occurring in the transfer; this quotient was then subtracted from one hundred. These percentage values are found in Table 15. For example the transfer from A to B reduced the number of retracing errors by 79.4%, and the cul de sac errors by 90.4%. Transfer in this case was more efficacious upon the cul de sac errors than upon those due to retracing. The purpose of this section is to determine whether transfer has a greater effect upon one type of error than upon the other.

Transfer on the whole exerts a slightly greater effect upon retracing. It tends to minimize the tendency to retrace relatively more than it does the tendency to enter cul de sacs.

The above conclusion is evident from an inspection of the data in Table 15. The percentage values are larger in 11 of the 16 cases of comparison. There is no essential difference between humans and animals as to the selective effect. Nor does the differential effect depend upon the direction of the transfer. Transfer is more effective upon retracing in some pairs of mazes than in others, but this difference of effect with the various pairs of mazes is not correlated with their similarity; neither is it correlated with the difficulty of the mazes measured by the number of errors involved in either original or transferred learning.

TABLE 15. GIVING THE PERCENTAGE OF DECREASE DUE TO TRANSFER IN NUMBER OF RETRACING AND CUL DE SAC ERRORS RESPECTIVELY.

RATS		
	Retracing	Cul de sac
A-B	79.4	90.4
A-C	38.0	61.5
A-D	78.8	82.2
A-E	65.4	41.9
A-F	40.2	45.0
B-A	96.8	94.3
C-A	64.9	61.7
D-A	87.9	60.4
E-A	64.3	57.2
F-A	63.8	43.8
HUMANS		
A-B	89.7	86.7
A-C	28.9	11.5
A-D	96.1	87.0
B-A	99.1	94.1
C-A	25.5	26.3
D-A	48.7	36.2

In those mazes in which the greatest saving of cul de sac errors is exhibited, transfer also tends to give the greatest saving in retracing.

The evidence to support this conclusion was determined by correlating the two sets of values given in Table 15. The results were .60 and .90 for the rats, and 1.00 and 1.00 for the humans. These correlation values mean that in measuring the degree of transfer between various pairs of mazes, one can utilize the saving in retracing, or the saving in cul de sac errors, or the saving in the total number of errors, as we have done, without materially changing the results.

F. SUMMARY.

Upon the basis of the foregoing study of transfer, we have been able to make the following conclusions.

I. The nature of the transfer is positive. The learning of one maze has a beneficial effect in the mastery of a subsequent maze situation. We tested the nature of the transfer for both

directions between five pairs of mazes with rats, and three pairs with human subjects. We thus secured sixteen separate tests of the nature of the transfer. In all of the sixteen cases the result was positive. We also used three criteria of measurement—trials, errors, time—thereby obtaining 48 separate measurements of the transfer. In 46 of these measurements the average for the test group was smaller than that for the control group, thus indicating a positive transfer. It might be argued that the difference between the original and transferred learning records was due to chance or group differences. That the differences were not due to chance, but to the positive transfer, was proven by the fact that the actual difference was found to exceed the probable difference in 45 of the 48 comparisons. Further, the consistency with which the differences occurred lessened the probability manyfold of chance being a primary causal factor. Our method of securing the subjects and of selecting the groups eliminated the probability of group differences functioning to a significant degree. Hence we conclude that the primary causal factor determining the differences between the original and transferred learning records is not chance or group differences, but the positive nature of the transfer.

In the 48 measurements of the transfer two exceptions occurred to the positive results. These were found with the human subjects by the criterion of trials. In these two instances we have shown that the loss was probably due to an individual peculiarity. Furthermore, in these two cases the transfer as measured by errors and time was of a strong positive character. While we admit the possibility that these two exceptions were caused by negative transfer, we believe that the greater probability favors the transfer being of a positive nature, and that we are justified in concluding that the transfer was positive in the sixteen tests.

The transfer remained positive despite our efforts to produce conditions that would give a negative result. In the first pair of mazes, A and B, the construction was so designed as to give a high degree of positive transfer, and the results verified our expectations. The purpose in designing the other four pairs of mazes was the desire to secure conditions that would produce

a negative effect, and in every instance we were disappointed in the outcome. We do not conclude that it is impossible to demonstrate a negative transfer effect between two pairs of mazes. On the contrary, we admit such a possibility. However, for the conditions maintained in this experiment, the transfer was of a positive nature.

2. Transfer is a composite process consisting of both positive and negative elements, and the total result is determined by the predominance of the one or the other of these elements. The total effect was positive although the presence of a negative element was shown to exist. Maze F was designed in relation to Maze A in such a manner that it was possible for us to determine whether certain habits acquired in A exerted a negative effect in the subsequent mastery of F. Subjects with the Maze A experience had greater difficulty in eliminating the tendency to enter section 6-10 in F, entered this section much more frequently, and made many more errors in this section, than did those subjects without such an experience. This evidence, we believe, proves the existence of a negative element in the transferred learning of Maze F. In order to produce a negative transfer effect, conditions will have to be arranged wherein the negative element predominates.

3. The degree of transfer is determined by a number of factors.

(1) It is in part a function of the nature of the second activity. The first activity was constant for all of the groups in the first experiment, while the second problem varied with each group. The divergence of the results was wide enough to indicate that the activity set up in the second maze situation was in part a causal factor. The proof to substantiate this conclusion was found in the fact that the relative amounts saved in the five cases was correlated with the relative difficulty of the second mazes as measured by the original records of mastery. A positive correlation was found in all six cases of comparison.

(2) The activities acquired in the first problem determine in part the degree of the transfer. In this instance the second maze was constant, and the varying activity was the first problem.

This conclusion was proven by the fact that the amounts saved and the original learning records were roughly proportionate, and by the consistency of the positive values found in determining the correlation between the degree of transfer and the difficulty of the first problem.

(3) The degree of transfer is dependent in part upon the degree of similarity of two maze patterns. The evidence for this conclusion was determined by computing the correlation between the degree of transfer the similarity of the five pairs of mazes. Two methods were utilized in securing these correlation values—the order of merit, and difficulty of mastery. This gave twelve values for each experiment. In the first case all twelve values were positive, while in the second 11 of the 12 values were positive. The significant thing here was the consistency with which positive values occurred. This fact enhanced their validity manifold, and, we believe, amply justified the above conclusion.

(4) The amount saved is determined in part by the direction of the transfer. Differences were found to exist for the opposite directions of transfer, stated in both absolute and relative terms. The influence of the direction of transfer in mediating the differential results was proven by the fact that the differences are a function of the degree of similarity of the various pairs of mazes. That pair of mazes most similar yielded the smallest difference of saving when the direction of transfer was reversed, while the largest difference of saving tended to obtain for the most dissimilar pair of mazes. This relationship was tested by computing the correlation between the size of the differences and the similarity of the various pairs of mazes. Seventeen values were obtained, and 14 of these were positive. We believe, that the consistency with which positive values occurred proves the above conclusion.

4. The laws and conditions of transfer are essentially identical for the two types of organism employed in this study.

The transfer was positive for both human and rat subjects. This was the result in the six cases of comparison when the maze patterns were identical, and the conditions of experimentation highly similar. The results were the same for the two types

of subjects in two respects. First: That pair of mazes which induced the greatest amount of saving for the rats had a like effect with human subjects; mazes which gave the least results with the human subjects produced a similar effect with the rats. This comparison was proven by determining the correlation values in absolute and relative terms. Of the nine values resulting seven were perfect and two were plus fifty. Second: The influence of the direction of the transfer was essentially identical for humans and rats. That direction which gave the greater saving for the rats also gave the larger values for the human subjects in six of the nine possible comparisons. When the correlations were computed in relative terms all of the values were perfect. While the total result was similar for rat and human subjects, there was one difference noted. The rats evidenced more ability in utilizing a previous maze experience in a new situation than did the human subjects. The difference favored the rats to a pronounced degree in all three pairs of mazes in both experiments when measured by trials and time, and whether stated in relative or absolute terms. Measured by the criterion of errors, the human subjects manifested the greater saving. A possible explanation of the differences was offered in the character of the previous training. Upon the total evidence, we concluded that human and animal organization are highly similar so far as the laws and conditions of transfer are concerned; that the processes involved were highly similar, and that no factors, such as rational activities peculiar to the human subjects, were functioning in the process of transfer.

5. A positive correlation was found between any two of the three criteria of measurement. There were 24 values in each experiment; in the first 23 of these were positive, and in the second all were positive. This indicates that some dependent relation exists between trials and errors, trials and time, and errors and time as a means of measuring the transfer. In the first experiment higher values predominated between time and errors, but in the second experiment there was little uniformity as to higher values. This lack of uniformity prevented us from drawing any conclusion as to the greater validity of any two of the three criteria of measurement.

6. The locus of transfer was on the average confined to the first five trials. The subjects were saved the equivalent of the first five trials of effort. Twelve of the sixteen comparisons approximated the average. The exceptions ranged in the amount saved from two to eight trials. This means that in the transferred learning the subjects attacked the new problem at an advanced stage, varying from the second to the eighth trial, and completed the mastery in a normal manner.

7. Transfer exerted some selective effect upon the types of error. The tendency to retrace the true pathway is minimized relatively more than the tendency to enter cul de sacs. This conclusion was supported by the fact that when the reduction in the two types of error was stated in relative terms, the values representing retracing were larger in 11 of the 16 cases of comparison. This differential effect does not seem to depend upon the direction of the transfer, the difficulty of the mazes, nor upon the degree of similarity of a pair of mazes. These facts indicate that the selective effect upon retracing represents some general transfer effect applicable to all maze conditions.

It was further proven that that pair of mazes which produced the greatest or the least effect upon retracing had a similar effect upon cul de sacs. This was determined by correlating the two sets of data, and high positive values resulted in every case. This means that the general effect of the transfer was the same for both types of error; that pair of mazes which gave the greatest reduction in retrace errors also produced the greatest effect upon cul de sac errors, but the effect was greater upon retracing. Thus we concluded that one could utilize the saving in retracing, cul de sac errors, or both without materially changing the results.

G. THEORETICAL DISCUSSION.

It is not our purpose to formulate a theory of transfer. We shall confine our discussion to a consideration of two of the prevalent theories in relation to our results.

Bagley¹ maintains that transfer depends upon the development of ideals. Habits of neatness acquired in one school subject

¹ W. C. Bagley, *The Educative Process*, p. 208.

will transfer to other school subjects only in so far as an ideal of neatness has been inculcated. An ideal in this sense implies some sort of ideational purpose, and to my mind the theory indicates the doctrine that transfer can occur only on an ideational level. This negative aspect of the theory is controverted by the data of this paper. Our facts prove rather conclusively that a high degree of positive transfer can occur on a purely sensori-motor level. Transfer was manifested by the rats, and it is generally presumed that such organisms do not possess ideational powers. Furthermore, the rats exhibited a greater degree of positive transfer than did the human subjects. If transfer is mediated only by ideational activities, we are forced to assume that human beings are inferior to rats in intellectual capacity, at least so far as the maze situation is concerned. That ideals were not an effective factor in our experiments is proven by the fact none of the human subjects were previously aware that they were to be transferred from one maze to another, while many of the subjects remained in entire ignorance of the fact that a new maze problem had been substituted in the course of the experiment.

Thorndike's theory² postulates that transfer occurs between two activities only when these activities possess identical neural elements or bonds of connection, and that the degree of transfer is proportional to the degree of identity of the bonds. This theory, as first stated, would not permit the existence of any negative transfer effect. The existence of negative transfer has been demonstrated by several experiments, and our results further prove that the total transfer effect between any complex set of activities is a composite affair consisting of both positive and negative elements. Poffenberger³ in a recent Columbia study, with the probable approval of Thorndike, has modified the theory so as to logically include the negative factor. He states that transfer occurs when there is at least a partial identity of bonds. When the bonds established in the first activity are broken in the acquisition of the second act, negative transfer results. When

² Thorndike; *op. cit.*

³ Poffenberger: *The Influence of Improvement in One Single Mental Process upon Other Related Processes.* Jr. Ed. Psych. Vol. 6.

the bonds remain unbroken and are utilized in the second activity, positive results are secured. This theory possesses a great deal of *a priori* plausibility. Our conclusion that the degree of transfer is in part a function of the degree of similarity of the two maze patterns would seem on first thought to confirm this theory. Our conclusion that the degree of transfer is dependent to some extent upon the direction of transfer would likewise militate against the theory, for logically the partial identity of neural bonds possessed by two sets of activities should be independent of their temporal order. However, we wish to maintain that neither of the above conclusions can be regarded as constituting either a refutation or a confirmation of the theory. It would be rather ridiculous to assert that any two sensori-motor activities must possess a degree of neural identity proportionate to the relative amount of effort expended in their acquisition. Theoretically one might acquire two activities totally isolated in the nervous system whose acquisition involved equal amounts of effort. Objective similarity of maze patterns does not necessarily mean a subjective similarity of the neural activities involved in their mastery. To my mind, herein lies the weakness of the Thorndike theory. Its validity can never be adequately tested. Any general agreement as to the degree of neural identity between any two complex problems is impossible. One can explain any fact of transfer on this basis, for all that is necessary is to assume the appropriate relationship between the nervous activities, and naturally it is practically impossible to disprove this particular assumption. Any theory or any explanation of the laws of transfer should possess some diagnostic value. One should be able to predict to some extent the degree of transfer to be obtained between any two activities. Thorndike's theory is defective in this respect.

Our facts indicate that transfer is to a large extent a function of the particular relationship existing between the two activities. A determination of the laws and conditions of this phenomenon must involve a thorough and complete analysis and definition of the essential relations obtaining between any two activities. This suggests that such complex activities as are involved in

the mastery of a maze situation constitute a poor medium for any comprehensive analysis of transfer. Experiments must be devised by the results of which we will be able to diagnose the relations between the two activities; such as keeping the reactions similar but varying the stimulus, or keeping the stimulus constant and varying the reactions. The problem is to simplify the situation and isolate and control the elements so as to ultimately analyze the existing relationships.

III. RETROACTION

A. RETROACTIVE EFFECT OF A CERTAIN ACTIVITY UPON VARIOUS OTHER ACTIVITIES.

The first experiment concerns the retroactive effect which the mastery of Maze A may exert upon the retention of Mazes B, C, D, E, or F. The questions at issue are the existence and nature of retroaction and its dependence upon the character of the habit affected. Two groups of subjects, a test and a control group, are necessary for each of the five pairs of mazes. In the test experiment, each of five groups of subjects learns one of the mazes B, C, D, E, and F; all are then transferred to Maze A; after thirty days mainly devoted to the mastery of A, each group is required to relearn its first maze. This procedure may be represented by the symbol B-A-B. In the control experiment, five groups are utilized and they repeat the above procedure with the exception of the mastery of Maze A; the symbol B—B thus represents the progression of events. The records secured in relearning Maze B in the test experiment represent the disintegration of the Maze B habits due to the thirty day interval plus the retroactive effect of the acquisition of A. The relearning records in the control experiment, however, represent merely the disintegration resulting from the thirty day interval. Any retroactive effect of A on B is thus measured by the difference in the amount of effort expended by the two groups in relearning B.

The test group on Maze B comprised 12 rats and 5 humans, while the control group consisted of 8 rats and 5 humans. On Maze C, both the test and the control group had 9 rats and 5 humans. The test group for Maze D consisted of 7 rats and 6 humans, and the control group had 6 rats and 5 humans. There were 9 and 7 rats respectively in the test and control groups for Maze E. The test group for Maze F consisted of 8 rats, and the control group of 7 rats.

TABLE 16. THE RETROACTIVE EFFECT OF A UPON B.

B-A-B						RATS		B—B					
Subj.	Trials		Errors		Time		Subj.	Trials		Errors		Time	
	L.	R.	L.	R.	L.	R.		L.	R.	L.	R.	L.	R.
1.	23	3	73	5	916	55	1.	33	0	113	0	1174	0
2.	58	1	460	1	12650	71	2.	49	0	189	0	1211	0
3.	94	4	409	6	7404	80	3.	53	23	205	17	1532	388
4.	40	0	265	0	2708	0	4.	43	8	128	9	1347	135
5.	69	0	238	0	2126	0	5.	73	0	304	0	2106	0
6.	35	0	96	0	1124	0	6.	51	12	204	23	1449	211
7.	49	30	234	37	1430	419	7.	59	12	198	9	1059	164
8.	44	43	191	48	1241	606	8.	48	0	155	0	990	0
9.	104	8	353	7	3204	142							
10.	45	11	209	12	1998	179							
11.	67	3	232	1	1764	57							
12.	88	39	224	36	1940	526							

HUMANS

1.	53	11	250	14	1669	322	1.	37	3	1003	13	1791	62
2.	30	0	190	0	903	0	2.	66	0	591	0	1780	0
3.	20	15	64	27	720	211	3.	6	0	36	0	598	0
4.	29	0	235	0	740	0	4.	9	2	116	3	500	29
5.	35	10	80	10	964	175	5.	50	0	287	0	1995	0

TABLE 17. RETROACTIVE EFFECT OF A UPON C.

C-A-C						RATS		C—C					
Subj.	Trials		Errors		Time		Subj.	Trials		Errors		Time	
	L.	R.	L.	R.	L.	R.		L.	R.	L.	R.	L.	R.
1.	30	1	414	71	12567	696	1.	78	0	167	0	1364	0
2.	57	0	202	0	1189	0	2.	78	1	286	7	2223	48
3.	35	1	188	1	1192	17	3.	78	0	246	0	2083	0
4.	36	3	177	7	1234	36	4.	78	0	281	0	1638	0
5.	74	0	250	0	1604	0	5.	78	1	329	1	4304	31
6.	58	3	475	8	3210	48	6.	46	2	175	23	6815	166
7.	12	2	169	2	264	26	7.	74	0	221	0	1647	0
8.	36	0	202	0	1041	0	8.	71	0	205	0	1907	0
9.	48	0	167	0	1310	0	9.	86	3	308	12	2502	128

HUMANS

1.	19	27	99	201	627	737	1.	25	12	174	10	774	114
2.	37	4	38	7	762	59	2.	25	0	121	0	864	0
3.	46	0	208	0	691	0	3.	48	12	188	13	479	13
4.	40	0	73	0	574	0	4.	67	3	147	1	792	13
5.	43	8	171	6	1110	89	5.	30	5	119	7	692	65

TABLE 18. RETROACTIVE EFFECT OF A UPON D.

D-A-D						RATS		D—D					
Subj.	Trials		Errors		Time		Subj.	Trials		Errors		Time	
	L.	R.	L.	R.	L.	R.		L.	R.	L.	R.	L.	R.
1.	13	0	145	0	2943	0	1.	7	0	130	0	1193	0
2.	23	6	122	5	1521	93	2.	19	3	91	7	4094	108
3.	8	0	203	0	3335	0	3.	38	3	144	5	1550	59
4.	31	1	164	1	1912	24	4.	12	5	73	30	1155	309
5.	25	3	394	25	8668	109	5.	11	1	175	5	3617	51
6.	10	4	109	18	2563	153	6.	11	0	163	0	1953	0
7.	9	0	84	0	1601	0							
HUMANS													
1.	13	15	84	53	600	380	1.	12	0	180	0	1158	0
2.	18	31	207	59	928	293	2.	3	4	67	22	416	328
3.	15	12	207	9	854	111	3.	3	0	5	0	70	0
4.	7	3	37	5	235	50	4.	5	4	90	3	466	51
5.	19	6	1113	17	2764	53	5.	31	0	202	0	903	0
6.	2	7	41	20	115	90							

TABLE 19. RETROACTIVE EFFECT OF A UPON E.

E-A-E					RATS		E—E						
Subj.	Trials		Errors		Time		Subj.	Trials		Errors		Time	
	L.	R.	L.	R.	L.	R.		L.	R.	L.	R.	L.	R.
1.	3	7	17	7	202	74	1.	8	0	36	0	277	0
2.	6	6	26	11	377	104	2.	1	0	14	0	105	0
3.	6	0	34	0	262	0	3.	6	0	6	0	312	0
4.	1	0	19	0	125	0	4.	3	0	7	0	208	0
5.	11	0	31	0	135	0	5.	2	4	7	5	178	50
6.	5	0	19	0	121	0	6.	1	4	14	2	127	26
7.	8	3	18	8	349	36	7.	5	0	20	0	221	0
8.	2	2	15	3	105	28							
9.	2	1	21	2	279	22							

TABLE 20. RETROACTIVE EFFECT OF A UPON F.

F-A-F					RATS		F—F						
Subj.	Trials		Errors		Time		Subj.	Trials		Errors		Time	
	L.	R.	L.	R.	L.	R.		L.	R.	L.	R.	L.	R.
1.	14	16	70	58	491	278	1.	17	1	86	10	816	37
2.	32	0	84	0	591	0	2.	21	5	163	9	1179	48
3.	23	25	159	177	1009	1499	3.	7	4	64	5	1688	27
4.	39	3	165	6	1090	38	4.	12	12	81	21	692	152
5.	35	7	186	86	1145	573	5.	30	5	115	28	909	135
6.	50	14	178	96	1371	704	6.	38	4	146	16	927	215
7.	39	20	139	103	1101	420	7.	34	13	105	16	1456	120
8.	28	26	186	51	3682	346							

The individual relearning records for the test and control groups and for each of the mazes are given in Tables 16 to 20 in the columns marked R. The differences between the two series of relearning records may be due to chance, group peculiarities, or retroaction. The possibility of group peculiarities being a large causal factor, as in the transfer experiment, has been obviated by our method of group selection. Any decision between chance and a retroactive effect presents difficulty because of the high degree of individual variability in the relearning records. Hence for the present we shall refrain from making any conclusions regarding retroaction and content ourselves with mere factual statements concerning the differences between the records of the two groups.

1. The occurrence of any disintegrating effect due to time, or to time and retroaction combined, is an individual matter.

Some individuals were affected and some were not susceptible to the influences. Individual exemptions were present in 14 of the 16 groups of subjects. Out of 82 rats, 31 were not affected, and 11 of the 31 human subjects manifested no disturbance. Of the total number of subjects employed in the experiment 37% suffered no disturbing effect. The human and animal subjects manifested no difference as to their immunity. If retroaction is present, its operations are certainly confined to specific individuals.

2. The individual variability in the relearning records is much greater than in the original mastery of the same maze.

For the purposes of this comparison we have inserted in Tables 16 to 20 the individual records for the original learning of the various mazes. These are found in the columns marked L. The values under L and R thus give the individual learning and relearning records respectively of the various subjects. The validity of our proposition is apparent from an inspection of these data. We computed, however, for each group the averages and the average deviations for the learning and relearning records. Each average deviation is divided by the corresponding average, thus expressing the value of the deviation relative to the average in percentage terms. These various percentile values are given

in Table 21. The columns L and R give the values for learning and relearning respectively. It is to be noted that the relative

TABLE 21. INDIVIDUAL VARIABILITY IN LEARNING AND RELEARNING.

RATS													
Group	Test Groups						Group	Control Groups					
	Trials		Errors		Time			Trials		Errors		Time	
	L	R	L	R	L	R		L	R	L	R	L	R
B-A-B	34	107	33	108	71	95	B-B	16	100	22	100	19	100
C-A-C	34	93	29	136	89	146	C-C	9	110	19	149	45	141
D-A-D	47	100	41	118	49	102	D-D	39	83	24	94	48	92
E-A-E	52	101	24	101	36	97	E-E	60	143	50	133	27	142
F-A-F	25	57	26	58	46	68	F-F	42	55	26	40	27	56
HUMANS													
B-A-B	26	84	45	81	27	80	B-B	62	120	77	122	46	120
C-A-C	17	113	44	157	22	133	C-C	39	70	16	73	15	70
D-A-D	42	58	91	78	53	71	D-D	80	120	60	136	57	133

variability in relearning is the greater in each of the 48 comparisons. On the average the relearning values are at least three times as large as those representing the original learning.

There is no consistent difference as to relearning variability between the human and rat subjects, nor between the control and test experiments. Comparing the variability of mazes, there is but one uniform result of whose validity we may be confident. Maze F gives by far the lowest relearning variability values for all three criteria and for both the test and control groups.

3. There is practically no correlation between the learning and the relearning records.

We ranked the individuals of each group according to their ability in learning as measured by each of the criteria. The same individuals were also ranked as to their ability manifested in relearning the same maze, and the correlation values between the two sets of data were computed. These values are to be found in Table 22. All of the values are extremely small, and but 28 of the 48 values are positive. If any tendency towards a positive correlation be present, it is slight in amount. No significant differences differentiate humans from rats, the control and test groups, nor one maze from another.

TABLE 22. CORRELATION BETWEEN LEARNING AND RELEARNING.

RATS			
Group	Test Groups		Time
	Trials	Errors	
B-A-B	.30	— .10	— .03
C-A-C	— .35	.44	.48
D-A-D	.50	.36	.22
E-A-E	.06	— .33	.55
F-A-F	— .37	.08	.41
Group	Control Groups		Time
	Trials	Errors	
B—B	.40	.39	.37
C—C	.33	.37	.85
D—D	.78	— .57	— .07
E—E	.05	.04	— .17
F—F	— .14	.16	— .46
HUMANS			
Group	Test Groups		Time
	Trials	Errors	
B-A-B	— .05	— .32	.38
C-A-C	— .25	— .42	.03
D-A-D	.26	.26	.15
Group	Control Groups		Time
	Trials	Errors	
B—B	— .10	.50	.00
C—C	.08	.70	— .60
D—D	— .22	— .20	— .40

This general result means that individuals making good records in mastering a maze are just as liable as not to do poorly in relearning the same maze. It is thus impossible to predict from the learning records the relative ability of a group of subjects in again mastering the same maze. The individual differences in the relearning tests thus do not represent any permanent differences of individual capacity, nor are they the result of any habits acquired during the original mastery of the maze. The greater individual variability manifested in relearning as opposed to learning (see former section) must thus be due, not to permanent individual peculiarities nor to acquired habits, but to individual differences in susceptibility to the disintegrating influences of intervening conditions. Since the relearning variability of the control group is not greater than that of the test group, it is evident that time and not any possible retroactive effect is primarily the responsible intervening factor.

4. A greater percentage of the subjects in the test series mani-

fested some degree of disintegration of the old habit. This greater effect may be due to retroaction.

This generalization is apparent from the data as tabulated in Table 23. Of the total number of rats used in the test series, 66.7% were disturbed while but 57% were affected in the control series. For the humans the percentages affected are 75 and 53.3 for the test and control series respectively. More rats were affected in the test series for 3 of the 5 mazes, and the same statement is applicable to the humans for 2 of the 3 mazes. On the whole there is a uniform tendency towards a greater susceptibility in the test series.

TABLE 23. PERCENTAGE OF SUBJECTS AFFECTED.

Group	Rats		Humans	
	Test	Control	Test	Control
B	75.0	50.0	60.0	40.0
C	55.5	44.4	60.0	80.0
D	57.0	66.7	100.0	40.0
E	55.5	28.5		
F	87.5	100.0		
Total	66.7	57.0	75.0	53.3

5. The average amount of disintegration for those subjects affected is on the whole somewhat higher in the test series.

The comparative data supporting the above generalization are to be found in Table 24. For the rats the test series gives the higher values in 13 of the 15 cases of comparison. With the humans higher values are found in the test series in all 9 instances of comparison. Considering the values for rats and for humans as single groups, the test series gives the higher values for all three criteria of measurement. The rat values in the test series are the larger in 3 of the 5 mazes, while the human values are higher for all three mazes. In the rat records the differences in favor of the test series are the most pronounced for Maze F. On the whole the humans show more evidence of negative retroaction than do the rats.

6. The percentage of disintegration or loss during the thirty day interval is on the whole somewhat higher in the test series.

Table 25 presents the comparative data in support of the above statement. The test series in the rat records gives the higher

TABLE 24. AVERAGE RELEARNING RECORDS FOR SUBJECTS MANIFESTING DISINTEGRATION.

				RATS			
Test				Control			
Group	Trials	Errors	Time	Group	Trials	Errors	Time
B	15.78	17.00	237.22	B	13.50	14.50	224.50
C	2.00	17.80	164.60	C	1.75	10.75	93.25
D	3.50	12.25	94.75	D	3.00	11.75	131.75
E	3.80	6.20	52.80	E	4.00	3.50	38.00
F	15.85	82.12	552.00	F	7.14	15.00	102.85
Average	8.18	27.07	220.27		5.88	11.10	118.07
HUMANS							
B	12.00	17.00	236.00	B	2.50	8.00	45.50
C	13.00	71.33	295.00	C	8.00	7.75	63.75
D	12.33	27.17	162.83	D	4.00	12.50	37.90
Average	12.44	38.50	231.27		4.83	7.41	49.05

values in 12 of the 15 cases of comparison, and in the human records higher values are found in all nine cases. Looking at the values for the rats as a single group, as shown by the averages in the table, the test series gives the higher values for 2 of the 3 criteria, while for the humans all three values are higher in the test series. For 3 of the 5 mazes the rat values are larger in the test series, while all three mazes give higher values for the humans. Maze D gives the greatest evidence of negative retroaction with humans, but with the rats this same maze shows more evidence of positive retroaction. Maze F in the rat records gives the greatest evidence in favor of negative retroaction. As was found in absolute terms, so also we find in relative terms that the humans manifest a larger disturbing effect of a retroactive character.

7. An analysis of the types of error gives evidence of a greater disturbing effect in the test groups.

Pechstein¹ has shown in his maze experiments that subjects learn the maze in about the same number of trials when retracing is prevented as when this type of activity is allowed. Retracing thus plays but little part in the mastery of a maze so far as the number of trials is concerned. It thus follows that the retracing present in the learning of a maze is an incidental and useless

¹ Pechstein. Op. cit.

TABLE 25. AVERAGE PERCENTAGE OF LOSS.

Group	RATS					
	Test			Control		
	Trials	Errors	Time	Trials	Errors	Time
B	20.8	6.3	10.2	13.2	3.7	8.2
C	4.0	2.8	2.4	1.1	3.0	1.1
D	11.5	4.0	2.0	12.7	9.3	5.8
E	57.9	17.5	12.2	85.0	12.4	7.0
F	50.2	49.9	44.6	36.2	13.5	11.4
Average	28.85	16.10	14.28	29.64	8.38	6.70
Group	HUMANS					
	Test			Control		
	Trials	Errors	Time	Trials	Errors	Time
B	25.0	12.2	13.2	6.0	0.8	2.0
C	34.0	45.0	26.6	19.0	3.8	7.8
D	132.0	27.5	35.5	42.6	7.2	18.0
Average	63.33	28.23	25.10	22.53	3.93	9.26

result of a peculiarity which manifests itself when an organism is in novel surroundings, or when it becomes lost or disturbed. The amount of retracing will in a measure represent the degree of disturbance due to novel conditions.

As a matter of fact, retracing was much more prevalent in the test experiments. Table 26 gives the data for the average number of retrace errors in the relearning of subjects manifesting a disturbance. The rats show a larger average number of retrace errors in the test group in all five cases of comparison, while in the human records the test group gives the higher averages in 2 of the 3 comparisons. Considering the rats and humans as single groups, we find that the test series has a greater number of retraces.

TABLE 26. AVERAGE OF RETRACE ERRORS IN THE RELEARNING OF SUBJECTS SHOWING A DISTURBING EFFECT.

Group	Rats		Humans	
	Test	Control	Test	Control
B	0.66	0.0	4.0	6.0
C	14.60	9.0	50.3	1.2
D	10.00	8.0	14.0	8.2
E	2.60	1.5		
F	26.30	4.1		
Average	10.83	4.52	22.76	5.13

Thus there appears to be a general tendency for more retracing to occur in the test series, which fact may argue for the presence of a disturbing or retroactive effect.

8. Certain peculiarities of behavior indicate a negative retro-active effect for Maze F.

By comparing the diagrams of Mazes A and F (Figs. 1 and 6), it is seen that the section numbered 6 to 10 as a cul de sac in Maze F, corresponds to an open section in Maze A, through which the rat must pass in his learning of this maze. After having mastered F, this additional route has to be added to his maze experience when transferred to Maze A; while in the relearning of Maze F this newly acquired habit must be again omitted. These conditions make it possible to determine in a fairly accurate manner whether or not the Maze A experience is functioning in the relearning of F. If the rats persist in entering the section 6-10, it is evident that the Maze A habit is functioning in such a manner as to interfere in again mastering F.

The record of the number of entrances into this section was kept for each subject. Rat A entered this section in 17 of his 22 trials; rat C, 13 out of 25 trials; rat E, 7 out of 7 trials; rat F, 8 out of 11 trials; rat G, 12 out of 20 trials; rat H, 4 out of 26 trials; while rats B and D made no entrances. The record for the control group is quite different. Five of the rats in this group entered this section only one time; one rat entered it twice, and one entered it three times. By this comparison, it appears that with the majority of the subjects in this group, the Maze A habit functioned in such a manner as to interfere in the relearning of Maze F.

9. The test series manifests the greater degree of imperfection of the maze habit on the first day's test.

Previously we have measured the degree of the imperfection of the habit by the time necessary to relearn the maze. This has been termed the relearning method. The usual method of measuring the imperfection of a habit in the experiments on retro-action is that of recall. In a maze experiment the nearest approach to the method of recall is to utilize the records of the activity of the first day in the relearning tests as a measure of the imperfection of the habit. We have made such a comparison, and the results are given in Table 27.

TABLE 27. FIRST DAY'S ACTIVITY IN RELEARNING.

Test			Control		
RATS					
	Errors	Time		Errors	Time
B-A-B	1.9	46.8	B-B	0.75	37.0
C-A-C	9.4	113.2	C-C	5.3	61.7
D-A-D	2.8	61.4	D-D	2.8	62.8
E-A-E	2.2	22.3	E-E	1.0	19.5
F-A-F	23.4	154.6	F-F	8.8	47.8
Av.	7.9	79.7	Av.	3.7	45.7
HUMANS					
B-A-B	1.2	31.6	B-B	2.2	46.5
C-A-C	8.0	31.8	C-C	1.0	17.8
D-A-D	7.7	32.5	D-D	3.2	57.4
Av.	5.6	31.9	Av.	2.1	40.6

On examining the data in this table, we find that with the rats the test group has the larger number of errors and the greater amount of time for four of the five mazes, while the human results give the larger number of errors in two of the three comparisons and the greater amount of time in one of the three cases. Considering all of the rats and humans as two separate groups, we notice that the average for the rats is larger for the test group for both errors and time, while for the humans the test group has the larger record only for errors. With rats 31% of the subjects in the test groups were perfect in the initial trials, while in the control groups only 30% gave perfect records. In the human control groups 53% of the subjects manifested no disturbance in the trials of the first day, while only 24% of the subjects in the test groups gave evidence of no loss. Thus there appears to be a general tendency for the test series to manifest a greater degree of imperfection on the first day of the relearning activity.

B. RETROACTIVE EFFECT OF VARIOUS ACTIVITIES UPON THE SAME PROCESS.

Our second experiment deals with the retroactive effect which Mazes B, C, D, E, and F may exert upon the retention of Maze A. The same questions are at issue, as in the previous experiment, and records from test and control groups are necessary to

determine these issues. Several groups of subjects mastered Maze A; one of these groups then learned Maze B, another Maze C, a third Maze D, one Maze E, and a fifth group was transferred to Maze F. After an interval of thirty days each group relearned Maze A. Another group mastered A and waited an equal length of time and relearned the same maze; this is the control group. The groups learning another maze in the interval are the test groups. The difference between the results from the test groups and the control groups is termed the retroactive effect of the second maze upon the retention of Maze A.

The control group for rats was composed of 11 subjects, and in the human control group there were 6 subjects. The test groups to determine the retroactive effect of B upon A consisted of 9 rats and 5 humans; to show the effect of C upon A 9 rats and 6 humans were used; in the test groups for D there were 6 rats and 3 humans; 8 rats comprised the test group for Maze E. The records from 9 rats are employed to determine the effect of F upon A.

The relearning records for all individuals in both the test and control groups are found in Tables 28 and 29. The symbols A-B-A, A-C-A, etc., and A—A indicate the test groups and control group respectively. Chance, retroaction, or group peculiarities may have caused the difference between the records of the two groups. The selection of the groups prevented group peculiarities from functioning sufficiently to cause the differences. Again we are confronted with such wide individual differences that we are unable to make any accurate judgment between chance and retroaction. In this section we shall also confine ourselves to dealing with the factual comparisons.

1. Any disintegration due to time, or time plus retroaction, is an individual matter.

Many individuals were affected and some manifested no disturbance in the tests for retention. Exemptions from the influences are found in 8 of the 10 groups of subjects. Thirteen out of 52 rats, and 6 out of 19 humans gave evidence of no disturbing effect. Of the total number of subjects employed only 73.3% were affected by time or time plus retroaction. Thus we

TABLE 28. RETROACTIVE EFFECT OF VARIOUS MAZES UPON A.

RATS													
A-B-A						A-C-A							
Subj.	Trials		Errors		Time		Subj.	Trials		Errors		Time	
	L.	R.	L.	R.	L.	R.		L.	R.	L.	R.	L.	R.
1.	40	8	191	21	2009	146	1.	18	1	138	3	1254	62
2.	43	2	231	3	3260	31	2.	25	0	160	0	751	0
3.	20	2	224	7	4874	189	3.	59	1	318	52	3176	595
4.	30	16	194	33	1601	428	4.	45	0	295	0	899	0
5.	47	2	270	7	4907	41	5.	28	6	199	11	1087	129
6.	48	0	173	0	1388	0	6.	52	3	254	62	1786	605
7.	34	1	177	20	1452	58	7.	14	0	109	0	2187	0
8.	40	0	250	0	2210	0	8.	30	11	216	10	872	162
9.	49	0	166	0	1353	0	9.	42	5	246	11	1419	149
A-D-A						A-E-A							
1.	76	2	299	2	2128	46	1.	35	4	150	3	2195	60
2.	31	12	184	27	1180	188	2.	47	3	162	5	2708	49
3.	27	7	144	10	922	84	3.	27	0	221	0	1084	0
4.	33	1	570	1	5112	25	4.	34	2	108	1	2166	53
5.	31	0	124	0	1384	0	5.	48	0	158	0	1213	0
6.	28	2	126	1	1582	76	6.	22	7	174	66	2465	596
							7.	26	2	136	9	573	51
							8.	44	5	169	4	1526	72
A-F-A						A—A							
1.	36	4	170	6	710	60	1.	73	5	293	9	1773	101
2.	37	2	196	7	1016	64	2.	68	3	281	2	2279	54
3.	43	3	211	3	1307	49	3.	52	2	242	4	1684	46
4.	26	10	143	25	893	176	4.	27	20	131	35	797	306
5.	22	0	139	0	498	0	5.	70	9	265	7	4602	137
6.	33	4	169	12	802	65	6.	43	0	193	0	1077	0
7.	23	5	172	19	2196	117	7.	44	0	230	0	1227	0
8.	28	9	98	20	587	146	8.	72	9	356	11	2502	149
9.	23	4	144	13	616	59	9.	69	0	307	0	2088	0
							10.	16	8	143	24	1466	140
							11.	61	1	294	8	2409	38

TABLE 29. RETROACTIVE EFFECT OF VARIOUS MAZES UPON A.

HUMANS													
A-B-A						A-C-A							
Subj.	Trials		Errors		Time		Subj.	Trials		Errors		Time	
	L.	R.	L.	R.	L.	R.		L.	R.	L.	R.	L.	R.
1.	29	45	135	53	706	491	1.	32	14	212	26	1036	144
2.	7	0	95	0	384	0	2.	24	32	197	75	607	311
3.	15	0	83	0	329	0	3.	2	11	48	18	441	257
4.	22	0	181	0	1702	0	4.	39	19	318	18	943	167
5.	23	5	180	53	510	182	5.	34	26	157	57	631	285
							6.	47	5	152	19	1471	135
	A-D-A							A—A					
1.	18	11	691	14	1851	134	1.	18	0	211	0	1005	0
2.	5	6	442	15	1585	105	2.	7	4	68	8	363	61
3.	40	38	360	71	1385	415	3.	23	6	226	16	1118	110
							4.	49	0	531	0	1157	0
							5.	21	0	193	0	1437	0

again make the statement that if retroaction is present, it is confined in its operation to certain individuals.

2. In comparing the individual variability in relearning with that in the original mastery of the same maze, we find much wider variations in the relearning.

The records for the original mastery, in addition to the relearning records, are inserted in Tables 28 and 29. The column marked L gives the values for the original learning, and the one marked R presents the individual relearning records for the various subjects. The above generalization is obvious from an examination of the comparative data in this table. In order to facilitate the comparison, we have expressed the value of the average deviation in relation to the average in percentage terms. These percentages will be found in Table 30. The columns L and R give the values for learning and relearning respectively. For the rats the relative variability in the relearning is greater in all 18 cases of comparison, while in the human records there is a greater variability manifested in the relearning in 10 of the 12 comparisons. On the average the relearning values are approximately twice as large as the values for the original mastery.

The two exceptions noted above are to be found in the human records; however, on the whole there is no consistent difference between the rat and human records. The two exceptions with the humans may be due to the fewer number of subjects in these two groups. No consistent difference is apparent between the test and control groups, nor between the various mazes.

TABLE 30. INDIVIDUAL VARIABILITY IN LEARNING AND RELEARNING.

Group	RATS						
	Trials		Errors		Time		
	L	R	L	R	L	R	
A-B-A	18	110	15	97	46	101	
A-C-A	37	97	29	108	42	92	
A-D-A	34	91	53	114	51	67	
A-E-A	23	65	14	125	37	110	
A-F-A	21	50	16	52	37	52	
A—A	29	87	22	85	36	81	
Group	HUMANS						
	A-B-A	35	140	27	119	53	72
	A-C-A	38	44	25	27	34	31
	A-D-A	85	72	41	75	27	65
	A—A	43	120	46	125	26	120

3. Our evidence fails to justify the conclusion that there is any correlation between the learning and the relearning records.

As in the former instance dealing with a similar correlation, we ranked the individuals according to their ability as manifested in the learning and relearning of the same maze. The correlation values between these two sets of data are given in Table 31. The majority of the values are too small to be significant, and only 16 of the 30 values are positive. From this it appears that there is no general tendency towards a positive correlation between the two sets of values. The differences between the humans and rats, the control and test groups, and the various mazes, are apparently not significant.

TABLE 31. CORRELATION BETWEEN LEARNING AND RELEARNING.

RATS			
Group	Test Groups		Time
	Trials	Errors	
A-B-A	— .37	.28	.33
A-C-A	.08	.56	.42
A-D-A	— .31	.39	— .60
A-E-A	— .21	.32	.45
A-F-A	— .14	— .25	.25
Control Group			
A—A	.13	— .04	.21
HUMANS			
	Test Groups		
	Trials	Errors	
A-B-A	.90	.38	.40
A-C-A	.00	.13	— .82
A-D-A	1.00	— 1.00	— .50
Control Group			
A—A	.00	.00	— .30

The result in this instance strengthens the validity of our conclusion in the corresponding section of the first experiment: The ability manifested by individuals in learning a maze is no index to their relative ability in relearning the same maze. Again we conclude that individual susceptibility to the disintegrating effects of time is responsible for the pronounced variability in the relearning records.

4. In the test series a greater percentage of the subjects gives evidence of some degree of disintegration of the old habit, than is found in the control series. Retroaction may be operating to cause this greater effect.

The data in support of the above statement will be found in Table 32. There was 76.08% of the total number of rats used in the test series disturbed by the influences, while 72.7% were affected in the control series. The human records show 80% and 40% affected in the test and the control series respectively. More of the rats were affected in 3 of the 5 test groups, while in 2 of the 3 test groups of the humans there was a larger percent affected. Two of the human test group had 100% of the subjects affected. A greater susceptibility to the influences is apparent in the test series.

TABLE 32. PERCENTAGE OF SUBJECTS AFFECTED.

Group	Rats		Humans	
	Test	Control	Test	Control
A-B-A	66.6	72.7	40	40
A-C-A	66.6		100	
A-D-A	83.3		100	
A-E-A	75.0			
A-F-A	88.9			
Average	76.08	72.7	80	40

5. For those subjects affected the average amount of disintegration is on the whole slightly higher in the test series.

For evidence in support of the above statement see Table 33. Higher values are found in the test series for the rats in 7 of the 15 cases of comparison. In the human records the test series give higher values in all 9 instances of comparison. Looking at the rat and human values as single groups, higher values are present with rats in two of the three criteria, with humans in all three criteria. No single maze in the rat records gives uniformly higher values in the test series, while all three mazes do so for the humans. For the rats Maze D gives the greatest evidence of positive retroaction, while Maze C manifests the greatest amount of disturbance. In the human records Maze B gives the greatest evidence for negative retroaction. The results indicate that the humans are more susceptible to negative retroaction than are the rats.

6. The disintegration or loss during the thirty day interval is on the whole somewhat higher in the test series when the results are stated in percentage terms.

TABLE 33. AVERAGE RELEARNING RECORDS FOR SUBJECTS MANIFESTING DISINTEGRATION.

RATS						
Group	Test			Control		
	Trials	Errors	Time	Trials	Errors	Time
A-B-A	5.17	15.17	151.00	7.09	12.50	121.38
A-C-A	4.50	22.83	263.67			
A-D-A	4.80	8.13	87.20			
A-E-A	3.83	14.67	146.83			
A-F-A	5.11	14.38	92.00			
Average	4.68	15.03	148.14	7.09	12.50	121.38
HUMANS						
A-B-A	25.00	53.00	336.50	5.00	12.00	85.50
A-C-A	17.83	35.50	216.50			
A-D-A	18.33	33.33	184.66			
Average	20.39	37.27	245.89	5.00	12.00	85.50

The comparative data upon which the above generalization is based are given in Table 34. The human test series give higher values in all nine cases of comparison, while the test series in the rat records give higher values in 7 of the 15 comparisons. Considering the average for the humans as a single group, higher values are found in the test series by all three criteria; the test series with the rats give higher values for 2 of the 3 criteria. Only Maze F gives uniformly higher values in the test series for the rats, and hence manifests a negative retroactive effect; Maze B evidences the greatest positive retroactive effect. All three mazes give uniformly higher results for the humans in the test series. The humans manifest a greater negative retroactive effect than do the rats.

TABLE 34. AVERAGE PERCENTAGE OF LOSS.

RATS						
Group	Test			Control		
	Trials	Errors	Time	Trials	Errors	Time
A-B-A	10.5	5.1	4.9	15.2	5.4	6.1
A-C-A	9.3	6.3	10.9			
A-D-A	13.3	4.2	5.8			
A-E-A	9.4	6.5	7.0			
A-F-A	16.0	8.3	9.8			
Average	11.7	6.08	7.68	15.2	5.4	6.1
HUMANS						
A-B-A	39.0	13.8	21.2	14.6	3.8	5.4
A-C-A	139.0	23.8	32.5			
A-D-A	92.0	8.3	14.7			
Average	90.0	15.3	22.8	14.6	3.8	5.4

7. A greater disturbing effect in the test series is evident from an analysis of the types of error.

In this connection we shall consider only the subjects that manifested some degree of disintegration. The comparative data for this topic are given in Table 35. The rats have a larger average number of retrace errors in the test series for 4 of the 5 mazes, and the human test series give the greater number of retraces in all three mazes. Considering the humans and rats as single groups, the test series for each group has the larger average number of retraces. Granting the hypothesis that the presence of retrace errors is an evidence of disturbance, we have here some evidence in favor of a retroactive effect.

TABLE 35. AVERAGE OF RETRACE ERRORS IN THE RELEARNING OF SUBJECTS MANIFESTING DISINTEGRATION.

Group	Rats		Humans	
	Test	Control	Test	Control
A-B-A	10.16	4.12	34.50	8.00
A-C-A	18.66		18.00	
A-D-A	0.00		8.66	
A-E-A	12.66			
A-F-A	4.87			
Average	9.27	4.12	20.38	8.00

8. A greater degree of imperfection of the maze habit is found with the test series on the first day of the test for retention.

In considering the retroactive effect in the preceding topics of this section, we have utilized the relearning or saving method. In this topic we wish to approximate the method of recall; hence we shall make use of the records of the activity of the first day in the relearning tests. The results for both the test and control groups will be found in Table 36.

By comparing the data in this table, we discover that for the rats the test group has the larger number of errors in four of the five cases of comparison, and the larger amount of time in three of the five cases. The test groups for the humans give larger values in the error column in all three comparisons, and for time they have the larger average in two of the three cases. The averages for all the members of the several test groups are larger than for those of the control group for the humans and

TABLE 36. FIRST DAY'S ACTIVITY IN RELEARNING.
RATS

	Test			Control	
	Errors	Time		Errors	Time
A-B-A	5.2	64.1	A-A	4.0	47.4
A-C-A	12.8	162.9			
A-D-A	2.5	43.2			
A-E-A	7.6	82.6			
A-F-A	4.4	37.2			
Av.	6.5	78.0			
HUMANS					
A-B-A	10.6	44.2	A-A	0.4	26.2
A-C-A	6.8	47.6			
A-D-A	4.8	24.8			
Av.	7.4	38.8			

the rats by both criteria. We believe that this evidence is sufficient to warrant the statement that there is a general tendency for the test groups to manifest a greater degree of imperfection on the first day of the relearning activity.

C. SUMMARY.

In treating the experimental results of our study of retroaction, we refrained from making any general conclusions. We have dealt thus far mainly with the factual material. We shall now summarize the results under two headings, and make our general conclusions.

I. Retention.

1. The groups exhibit a wider range of individual variability in relearning a maze than in its original mastery. But two exceptions to this generalization were found in 78 cases of comparison. This greater variability is not a function of the maze, the kind of subject, nor is it dependent upon the type of activity interpolated between the learning and relearning. It is primarily a function of the individual's susceptibility to the disintegrating effect of time.

2. There is no correlation between the learning and relearning records. The majority of the correlation values are too small to be significant; neither are they consistently positive or negative. Of 78 cases, 44 were positive, 30 negative and 4 cases

showed zero values. Subjects manifesting good ability in mastering a maze may thus do poorly in relearning the same maze. The individual differences either in mastering or relearning a maze are thus due to chance rather than representing individual differences in ability. Neither is there any correlation between learning records and retentive capacity. Both quick and slow learning groups may exhibit the maximum of retentive capacity, and both classes may be equally susceptible to the disintegrating effects of time. This relation between acquisition and retentive capacity is not dependent upon the type of subject, nor upon the character of the intervening activity.

3. Human subjects manifested a slightly greater retentive capacity of sensori-motor habits than did the rats. This conclusion is based upon a comparison of the relearning records of rat and human subjects for the same mazes, viz. A, B, C, and D. Only those subjects employed to test the disintegrating effect of time were used in this comparison.

A greater percent of rat subjects showed some disintegration in three of the four maze comparisons. For Maze C a larger percentage of the human subjects evidenced some degree of disintegration. Of the subjects that forgot part of the habit during the interval, the average relearning records for all three criteria of measurement are larger for the rats for Mazes A and B; the same condition prevails for Maze D for errors and time, and for Maze C for time alone. Thus the rats have the larger averages in 9 of the 12 comparisons. Considering the rats and humans as single groups, the averages for the rats are larger by all three criteria. The average percentage of loss is on the whole larger for the animals. The rats have the larger values by the three criteria for A and B, and by errors for Maze D. The humans manifested a higher percentage of loss by all three criteria on Maze C, and on Maze D by trials and time. The average of the percentage values for the rats and humans as single groups is the larger for the animals by errors only, the humans having larger values for trials and time. This result is perhaps due to the fact that the human group on Maze D exhibited a very high percentage of loss by trials and time, due

to an individual peculiarity. In the first day's activity in the tests for retention, the rats evidenced the greater amount of disturbance. The averages for the rats are the larger in 5 of the 8 comparisons. As single groups the rats have the larger values for both errors and time.

4. The degree of retention is a function of the maze activity. In comparing the different mazes in respect to retention, we find that for human subjects Maze D gives the largest average percentage of loss, and Maze B the smallest percentage of loss by all three criteria. A similar condition is noticed to exist with rat subjects. To test whether this is a real difference or a mere matter of chance, we figured the correlation between the percentage of loss due to time and the difficulty of mastery. The mazes were arranged in order of their greatest difficulty, and again arranged in their order of greatest percentage of loss after a thirty day interval. The values for the rats are $-.542$, $-.922$, and $-.714$; and for the humans $-.400$, $-.725$, and $-.800$. Thus all of the six values are seen to be negative. This means that the maze which required the greatest effort to master was retained better, and that maze requiring the least effort to master manifested the greatest percentage of loss; in other words, this means that the deeper a motor habit is driven in by prolonged effort, the longer and better it will be retained.

5. A positive correlation exists between any two of the three criteria used in measuring retentive capacity.

The subjects in each group were ranked from the lowest to the highest by each of the criteria, and the correlation was computed between trials and errors, trials and time, and errors and time. This procedure gives us 30 correlation values, all of which are positive. Of the thirty values, 9 were perfect and 15 were above .90; four were between .70 and .80, one .52, and one .38. The degree of correlation is not a function of the maze, nor of the type of subject employed. The relation between any one pair of criteria is no more significant than that existing between any other pair. These results indicate that we can utilize one, any two, or all three criteria in the measurement of retentive capacity and obtain practically the same results.

II. Retroaction.

I. The greater degree of disintegration occurred for the test groups. This result indicates that negative retroaction was present in our experiments. The difference is not great, but it is consistent for all rubrics of comparison for both rats and humans, and in both experiments. The validity of the conclusion must be based primarily upon the consistency of the results. The acquisition of any maze activity must thus be regarded as exerting some disintegrating effect upon maze habits previously mastered. The evidence in support of the above conclusion follows.

(a) On the average a greater percent of subjects was affected in the test series. This result obtained for both humans and rats in both experiments. The test series manifested the larger percentage in 3 of the 5 mazes for rats in both experiments, and in 2 of the 3 mazes for humans in each of the experiments.

(b) Limiting the comparison to those subjects affected, the rats exhibited the greater disintegration in the test series stated in absolute terms, in 5 of the 6 comparisons. The humans also exhibited the greater disintegration by all three criteria of measurement for both experiments. Considering the mazes as units, the humans gave the poorer records in the test series in each of the 18 instances of comparison. Likewise, the rats made poorer records in the test series in 20 of the 30 comparisons.

(c) The test series gave the poorer records in both experiments when the results are stated in relative, or percentage terms; that is, the test series exhibited the greater percentages of loss. The rats as a group gave the poorer average records for both errors and time in each of the experiments. The records for the humans were the poorer by all three criteria for both experiments. Considering the mazes as units, the humans gave the poorer records in each of the 18 instances of comparison, and the rats for 19 of the 30 cases.

(d) The average amount of retracing was the greater in the test series for both humans and rats in both experiments. The retracing for the rats was the greater for 9 of the 10 mazes, and for 5 of the 6 mazes with the human subjects.

20

(e) The greater disintegration occurred in the test series when measured by the records of the initial trials. This result obtained for the rats as a whole in 4 of the 5 instances of comparison, and for the humans in 2 of the 3 comparisons, in both experiments; it was likewise true for the rats in 9 of the 10 mazes, and for the humans in 4 of the 6 mazes.

2. The existence of retroaction is a function of the individual. In practically every group in both experiments, some subjects were affected and some were not. Of the total number of subjects employed in the first experiment, 37% manifested no disintegration, and in the second experiment 26.7% of the subjects gave no evidence of a disturbing effect.

3. Human subjects are more susceptible than rats to the disintegrating effect of retroactive influences.

For purposes of this comparison, we shall utilize only the records for Mazes A, B, C, and D upon which both rat and human subjects were employed. This gives us six cases of comparison, three in each experiment. To determine the amount of retroaction present in each test, we subtracted the records of the control group from those of the test group; the difference may be said to be due to retroaction. Upon a basis of a comparison of these results, we have the following evidence in support of the above conclusion.

A larger percentage of the human subjects manifested some degree of negative retroaction. The percentages are considerably larger for the humans in 4 of the 6 cases. The two exceptions are found in noting the effect of B upon A and of A upon B; the relative susceptibility of the two groups may thus be a function of the maze. In three of the human groups 100% of the subjects were affected, while in no instance did any of the rat groups have all of the subjects affected.

Stated in absolute terms, the average amount of disturbance for the human subjects is the larger in all six of the mazes. In the 18 instances of comparison, the values are the larger in the human records in 17 cases. Taking the groups as units in the first and second experiments, the humans have the larger averages in all six comparisons.

Making the comparison in terms of percentage of loss, we find that the humans manifested a greater degree of disintegration in all six mazes, and by all 18 instances of measurement.

The human subjects manifested the greater disturbance as measured by the amount of retracing. The average for the humans is larger in four of the six mazes. Considering the subjects in each experiment as units, the averages are much larger in both cases for the human subjects.

In the first day's activity of the relearning tests, the humans evidenced a greater disturbance in 4 of the 6 mazes when measured by errors, and in 2 of the 6 mazes when measured by time. The result by time may readily be attributed to the fact that the rats in returning to a maze problem after an interval of several days are quite cautious at first and travel rather slowly.

4. A positive correlation was found to exist between any two of the criteria employed in measuring the retroactive effect. The ranking method was used, and the values were determined for trials and errors, trials and time, and errors and time. In this manner we secured 12 values evenly divided between the rats and humans. Nine of the 12 values are positive. The validity of the conclusion depends upon the consistency of the positive values. This result is not a function of the maze, nor the type of subject. A much closer and more significant relation exists between errors and time than between any other pair. Of the four values measuring this relation, three are perfect and one .70, while none of the other values are above .50. In measuring the retroactive effect of one maze activity upon another, we can thus utilize one, two, or three of the criteria without materially changing the results.

5. Direction is not a deciding factor in determining the amount of retroaction present. The manner in which our experiment was arranged makes it possible to determine whether the retroactive effect of two activities upon each other was the same or different; thus we have the retroactive effect of A upon B, and of B upon A, and so with the other pairs of mazes. No new data are needed; a comparison of the data employed in determining the amount of retroaction present supports the above

conclusion. When the mazes are ranked according to the amount of retroaction present, we find that in both directions with rat subjects A and F stand first, and A and D last. Each of the other three pair has a slightly different rank according to direction. In the human records we have a similar result. The largest disturbing effect occurred between A and C in both directions, while the rank in the other two pairs of mazes differs slightly for the two directions. As a matter of chance, we would expect some difference due to the change of direction. We believe that the fact that three pairs of mazes are not affected by the change of direction is more significant than the difference in the other pairs, and that the above conclusion is justified.

6. The degree of retroaction is a function of the interpolated maze activity. The easier is the maze to learn, the greater is the resulting negative retroaction. The various mazes were ranked in their order of difficulty of mastery. These were also ranked in order as to their retroactive effect. Correlation values between the two values were computed. Five of the six values are negative, three of which are perfect and two above .60.

A positive retroactive effect was secured for Mazes A and D with rat subjects. The interpolation of A was beneficial upon the remastery of D by all three criteria. Likewise D exerted a favorable effect upon the relearning of A by all three criteria. This fact indicates that retroaction may be positive in character with some pairs of mazes. Additional evidence in support of the above conclusion is had as a result of computing the correlation between transfer and retroaction. These values will be discussed in the following topic.

7. There is a negative correlation between positive transfer and negative retroaction. Those conditions which produce the maximum amount of positive transfer give the least amount of negative retroaction. We ranked the mazes according to the percent of transfer present in both of the experiments dealing with that topic. We also ranked the mazes according to the amount of negative retroaction present determined by subtracting the records of the control groups from those of the test groups. From these ranks we computed the correlation between mazes;

in the first instance the correlation is between *A-B* etc. in the transfer and *A-B* etc. in the tests for retroaction, while in the other case the values are between the records of *A-B* etc. in the transfer and *B-A* etc. in the retroaction. This was done for both experiments, thus giving us 24 values in all. Twenty-one of the values are negative, 5 of which are perfect, 7 are $-.70$ or above, and the remaining values are around $-.50$. These results mean that the greater the assistance rendered by the first maze experience in the mastery of the second maze, the less disturbance will there be in relearning the first maze.

The same relation between retroaction and transfer is apparent from a comparison of the rat and human subjects. The rats manifested a greater ability than did the humans in the transfer experiments, while they exhibited the lesser susceptibility to retroactive disturbances.

D. THEORETICAL DISCUSSION.

Two explanatory conceptions of retroaction which possess some degree of logical plausibility may be suggested.

1. The Transfer Hypothesis. This conception has been suggested by DeCamp. The retroactive effect is regarded as a case of transfer. In the maze sequence *A-B-A*, the term retroaction refers to the effect of the acquisition of the *B* habit upon the subsequent functioning, or relearning of the *A* habit. The transfer hypothesis assumes that this effect is mediated by the simple transference of certain elements of the *B* habit to the succeeding maze *A* situation. Theoretically this transference may operate either in an advantageous or detrimental manner; in other words retroaction may be positive or negative.

2. The Disruption Hypothesis. In the maze sequence of *A-B-A*, we know that transfer obtained in proceeding from *A* to *B*. Certain elements of the complex *A* habit have been transferred to and utilized in the maze *B* situation. The hypothesis assumes that this incorporation of certain components of the *A* habit into the subsequently acquired *B* habit must necessarily involve its partial disruption and disorganization. The *A* habit is a complex system whose component elements have been

welded and associated into a unitary whole. The utilization of certain of its parts in a new situation must involve their dissociation from their former contextual relations, and the habit must thus be partially disrupted and disorganized. The remastery of A after the acquisition of B must repair not only the ravages due to time, but dissociate these elements from their new context and weld them anew into their original system of relations. According to the disruption hypothesis, the retroactive effect will invariably be negative in character.

The two hypotheses are not antagonistic or mutually exclusive. They may supplement each other. The effect of the B habit upon the functioning, or remastery of A may be due in part to the process of disruption, and in part to the transference of certain components of the B habit which are carried over to the subsequent A situation.

Several of our factual data are relevant to a consideration of the validity of the disruption hypothesis.

1. Transfer was present for all pairs of mazes employed in our experiments, and some degree of retroaction was manifested for each of these maze situations. This fact supports the hypothesis, for it assumes that retroaction is a necessary by-product of the previous transfer process.

2. All subjects, both human and animal, manifested the phenomenon of transfer, yet a certain percentage of these individuals (33) were not subject to any retroactive effect. This fact constitutes a serious objection to the acceptance of the hypothesis. These individual exceptions can be explained by the supposition that the retroactive effect is the combined result of both disruption and positive transfer, and that these two antagonistic tendencies were equal in these particular cases.

3. According to the hypothesis, retroaction will invariably be negative in character. Our results support this assumption in fourteen of the sixteen pairs of mazes. The pair A-D constitutes a possible exception with the animal subjects. We have previously indicated that there are good reasons for concluding that the retroactive effect was positive for this pair of mazes.

4. Transfer and retroaction are inversely correlated; those

conditions which give the greatest amount of positive transfer give the least subsequent negative retroactive effect. This fact presents some difficulty to the hypothesis, for it is logical to suppose that the degree of disruption will vary directly with the degree of transfer. This apparent exception to the validity of the theory may be obviated by the following mode of explanation. Both positive and negative transfer will produce a disruption of the first habit, and thus mediate a negative retroactive effect. We may assume that the disrupting effect of negative transfer will be greater than in the case of positive transfer. Our experiments on the transfer phenomenon have demonstrated that the total transfer effect is the sum of both positive and negative elements, although the positive factor predominated in every case. This conception explains the varying degree of transfer for the different conditions under which the experiments were conducted. The total effect is minimal when the negative factors approximate the positive in strength. The greatest total effect is achieved when the relative functional efficiency of the positive factors is at a maximum. Granted that the negative factors possess the greater disrupting efficacy, it is thus possible for the degree of disruption, and hence for the degree of retroaction, to be inversely related to the degree of transfer.

5. The retroactive effect was manifested on the first day's trials. This initial inefficiency of the A habit in the test group as compared with the control group indicates the presence of some previous disrupting or disorganizing process. The fact can, however, be easily explained in terms of the transfer hypothesis.

The following facts are pertinent to a consideration of the validity of the transfer hypothesis concerning the nature of the retroactive effect.

1. The retroactive effect was negative in fourteen of the sixteen pairs of mazes. If this effect is mediated by a transfer process, we are forced to conclude that this transfer was negative in the majority of cases. The same pairs of mazes were utilized in the transfer experiments and the effect was positive in every case. In other words, the hypothesis assumes that

transfer is negative when previous experiments on the same mazes have demonstrated that the effect is positive.

- A certain percentage of subjects manifested no retroactive disturbance, and yet all subjects exhibited transfer for the same pairs of mazes. The hypothesis is thus forced to reconcile the occasional absence of transfer in the one experiment with its invariable presence in the other, the same pairs of mazes being used in both cases.

The above facts, however, do not disprove the transfer conception of the retroactive effect, for the two experiments differed in several important respects, although the same pairs of mazes were employed in both. a) In the transfer experiments we were concerned with the effect of the first habit upon the acquisition of a second activity. The retroactive experiment was concerned with the effect of a second habit upon a third activity. In a series of successive activities, it is possible that positive transfer may obtain for the first pair while a negative effect will be exhibited by all succeeding pairs of the sequence. Some recent experiments in this laboratory, however, have disproved this assumption; some degree of positive transfer was invariably obtained in a sequence of five maze activities. b) The transfer experiment was concerned with the effect of a maze habit upon the *original mastery* of a second maze; retroaction refers to the effect of a maze habit upon the *remastery* of a second maze. In the one case, we are concerned with the utilization of a habit in the development of a *new* habit, and in the other with its utilization in the perfection of an *old* habit. The character of the transfer process may differ radically in the two cases. There are two considerations which support this assumption. (1) In the transfer experiments the subjects first develop the maze A habit. They are now transferred to a similar situation, viz., Maze B. This new sensory situation, because of its similarity, arouses or stimulates certain components of the previous A habit. The conditions of the retroactive experiment are radically different. The subjects have acquired two habits, A and B, in succession. They are now transferred back to the old Maze A. This sensory situation now tends to rearouse both the A and B habits.

Conflict and interference between the two systems must be the logical result. The conflict may apply to the process of recall, the transfer of the B habit operating to repress or to prevent the rearousal of the A habit which is to be remastered. Certain components of both systems may be reinstated and the conflict will result from their functional antagonism. In either case, confusion and an increased difficulty of mastery will result. In other words, transfer will be detrimental in the remastery of a maze although beneficial in its original mastery. (2) The transfer experiment demonstrated that the transfer effect was limited almost exclusively to the early stages in the development of a habit. The process of transference starts the subjects at an advanced stage of the problem and exerts relatively little effect upon the final development of the habit. In other words, transference exerts radically different effects upon the initial and the final stages in the development of a habit. In the retroactive experiment we are dealing with the remastery of a partly disintegrated habit. The habit is largely retained, the subjects are introduced to the problem at an advanced stage of mastery, and the process of relearning is essentially similar to the final stages of the development of a new habit. This fact is evident from a comparison of the curves of learning with those of relearning in the tests for retention. The relearning curves do not exhibit the pronounced initial descent characteristic of the typical learning curve; they approximate in character the latter part of the normal curve. Since transference exerts radically different effects upon the initial and the final stages of learning, one can not assume that transfer must produce the same effects in our two experiments; in fact one must assume that the effects are essentially different.

2. The transfer experiment does prove, however, that transference of some sort is always in evidence throughout a sequence of maze activities, and necessitates the conclusion that transfer of some kind did occur in the retroactive experiments, although it will not justify any assumptions as to the nature and degree of these effects.

3. The existence of a transfer process in the retroactive experiment is proven beyond doubt by certain facts which have been described on page 68. As previously noted, Mazes A and F were so designed in relation to each other that any transference from one to the other is easily detected. Both possess a common section 6-10. This section constitutes a part of the true pathway in A, but it is one of the cul de sacs in F. In the retroactive sequence of F-A-F, the subjects are first required to avoid this section, then to develop the habit of entering it in A, and then to avoid it again in the mastery of F. If the habit of entering this section acquired during the mastery of A is transferred to the subsequent F situation, it will be manifested by a greater number of entrances into this section than are made by the control group that has been given the maze sequence F—F. Such results were, in fact, secured. The average number of entrances into this section by the test and control groups respectively was 11.25 and 1.71. The corresponding number of trials in which this section was entered was 7.62 and 1.42. The number of trials necessary to eliminate this tendency was 10.35 and .085 for the test and control groups respectively. The total error scores made in this section for the two groups were 35.37 and 5.28. All of the above values are also much larger for the test group when stated in percentage terms.

These facts prove not only that a transference process does exist in the retroactive experiment but also that it was negative in character for the maze sequence F-A-F. It is also well to note that the greatest negative retroactive effect was secured for this particular pair of mazes.

4. The negative retroactive effect was evident in the first day's trials. This fact can be explained by the hypothesis. Our experiment demonstrated that the process of transference was more effective upon the initial trials than upon the later stages. If the negative effect is due to the confusion or conflict between the two systems of habits, this confusion will necessarily be present at first.

To summarize: The existence of a negative transference process has been demonstrated for one pair of mazes in the

retroactive experiment; the situations in our two experiments are so different that one is justified in assuming (unless proof to the contrary is advanced) that both sets of results were mediated by transfer; all of our factual data concerning the phenomenon of retroaction are readily explicable in terms of the transfer conception. On the other hand, there is no positive proof of the existence of a disrupting process. The conception explains some facts quite readily, some with difficulty, while others are incapable of explanation in such terms. We are thus forced to conclude that the retroactive effect is to be explained mainly, if not wholly, in terms of transfer, although it may be due in part to a process of disruption. It is well to remark that retroaction in so far as it can be reduced to transfer can not be regarded as a phenomenon *sui generis*.

Psychological Monographs

EDITED BY

JAMES ROWLAND ANGELL, UNIVERSITY OF CHICAGO
HOWARD C. WARREN, PRINCETON UNIVERSITY (*Review*)
JOHN B. WATSON, JOHNS HOPKINS UNIVERSITY (*J. of Exp. Psych.*)
SHEPHERD I. FRANZ, GOVT. HOSP. FOR INSANE (*Bulletin*) and
MADISON BENTLEY, UNIVERSITY OF ILLINOIS (*Index*)

STUDIES FROM THE PSYCHOLOGICAL LABORA-
TORY OF THE UNIVERSITY OF CHICAGO

The Reliability of Mental Tests in the Division of an Academic Group

By

BEARDSLEY RUML, PH.D.

Instructor in Psychology
Carnegie Institute of Technology

PSYCHOLOGICAL REVIEW COMPANY

PRINCETON, N. J.

AND LANCASTER, PA.

AGENTS: G. E. STECHERT & CO., LONDON (2 Star Yard, Carey St., W. C.)
LEIPZIG (Koenigstr., 37); PARIS (16 rue de Condé)



ACKNOWLEDGMENTS

The writer is indebted to Professors James R. Angell, Harvey Carr, J. W. Hayes, H. O. Rugg, A. C. Lunn, Dr. H. D. Kitson, and Dr. S. B. Vincent for helpful suggestions and criticisms given during the course of this research, to Professor L. C. Marshall and Dr. H. D. Kitson for assistance in securing data, and to Professor W. V. Bingham for the writer's interest in the field of mental tests.

I

The object of this research is the determination of the value of mental tests in dividing large groups of students into smaller groups of relatively equal mental ability. Such a division of students may be desirable for two reasons; the groups may be merely indicated in order that administrative officers may direct and advise students with greater confidence, or the large mixed class may be actually broken into groups of greater intellectual homogeneity so that no class need contain individuals who differ greatly in intelligence. In institutions where several sections of the same course must be offered, the sections might be formed on the basis of the abilities of the students instead of by a division of the total group according to the first letter of the individuals' surnames.

To indicate more clearly the sort of heterogeneity in which we are interested, it may be desirable to mention several ways in which students may differ from one another. In the first place, individuals may differ in the amount and quality of their previous academic training. The school attempts to remove these differences by imposing entrance requirements, and by providing courses in many departments suited to various degrees of preparation. Secondly there are differences in interests. That most institutions do not consider these differences of great importance in the first year of college is seen in the fact that the work of the freshman year is largely prescribed. Allowance is made for differences in interests in a wealth of elective courses in the junior and senior years. Thirdly, individuals may vary in *persistence* independently of differences in interests. Persistence of motive, the common character factor for which Webb²⁰ argues, may be the basis for one type of individual variation. In the fourth place there may be differences in special abilities. It is not possible at present to estimate the importance of special ability *independent* of general ability and special interests; but the high average correlation found by Carey⁴ between geography, science,

history and mathematics, $+.74$, indicates that special ability in any one of these subjects cannot be so important a factor as is popularly supposed. "The correlations between the various college subjects are all positive and argue against the commonly expressed belief in rather close specialization of abilities; the student who does well in one of these subjects tends to do well in all of them."¹⁰ Occasional cases of rare special ability in particular college subjects will doubtless occur, but such cases are easily recognized.

Lastly, individuals may differ in general intelligence, or brightness. The remainder of this monograph will attempt to show that by the use of mental tests, classes of students relatively homogeneous in this respect may actually be selected from the mixed group. It will further discuss from the standpoint of accuracy the advantages of different divisions of the total group. Clearly the total group may be divided into any number of sub-groups, and these sub-groups may contain any percentage of the total group. The problem is to choose the most accurate method of division. And finally, some evidence of a novel sort will be presented showing the relation between performance in mental tests and ability.

II

I

The discussion of the mental tests to be used in the selection of homogeneous classes, and of the methods of treating the measurements must be postponed pending the selection of a criterion by which the reliability of such tests may be judged. There is no a priori reason for believing that a series of mental tests will give an accurate indication of the intelligence of the individual tested. It is not hard to understand why such a series of tests *might* give a good index of ability, but to assert that this is the fact would be sheer dogmatism. As a matter of fact, it is seriously questioned whether tests devised to measure certain particular functions, such as memory or attention, really give a fair representation of those activities, so that considerably more doubt may

be attached to the value of psychological tests as measures of the less well defined function, general intelligence. It seems imperative, therefore, to abandon once and for all, performance in all the tests combined as the basis for judging the accuracy of any one test, not because standing in all tests combined is an unreliable measure of intellectual ability, but because it is not known in advance to be actually reliable.

Since the mental tests are to be used in a practical situation, in the division of a mixed group into smaller groups of more homogeneous intellectual ability, it seems natural to choose as a criterion of their value the success with which they handle the concrete problem of making the separation. The evaluation of the tests then consists in giving them experimentally to several groups of freshmen, and in determining how accurately the group would have been divided had the tests actually been in use.

The problem of evaluating the tests with reference to this concrete situation is therefore seen to presuppose the solution of another problem, namely that of determining how the experimental groups which have been tested should have been divided. If there were sharp and discrete classes of intelligence, for example, *good*, *mediocre* and *poor*, in which individual abilities might be placed, there would be little difficulty in finding how the separation should have been made. But unfortunately, general intelligence does not lend itself to any such rigid classification. On the contrary it seems to proceed through a continuous range of variation from the very bright to the very dull. As a result, a relatively homogeneous class would consist of the best or worst ten, twenty or thirty percent of the entire group. Such considerations show that it is impossible to designate any *definite* percentage of individuals as bright or dull. And, therefore, before any division whatever can be made, the individuals in the experimental group must be arranged in the order of their abilities. In other words, since general intelligence does not naturally divide into clear-cut categories, it is necessary that a value of some sort be predicated of the ability of every individual.

The mental ability of the individuals in experimental groups may be approximated in several ways. We have attempted to

show that to use as a measure of this ability the performance in the combined tests is inconsistent when the tests themselves are the objects of investigation. There remains the possibility of using either the estimations of instructors or other indices of the actual achievements of the students, provided it can be shown that either of these measures will properly evaluate the abilities of the students.

Two limitations to the usefulness of the tests seem to follow from selecting either estimations or achievements as the criterion by which the reliability of the tests in our experimental groups is measured. Since the tests are evaluated high or low according to the resemblance between the orders of individual abilities as given by the tests and as given by the criterion, it seems logical to inquire at what point the need for tests arises; why not use the information given by whatever criterion is adopted, as the basis for securing homogeneity? It may be urged that if the tests are required only to duplicate the information given by some other method already available, the tests in reality add nothing and had best be neglected as an unnecessary encumbrance to an overfull administrative program.

There are two answers to this objection. If the resemblance between the order of individuals as given by tests and as given by the criterion were complete, then the possibility that the tests might give new information would be eliminated. But the following facts must be taken into consideration: no matter how carefully the criterion of ability is selected, the chances are that it will be slightly in error. Now if the tests give a fairly high approximation, but not a duplication, of the order given by the criterion, there is reason to suppose that in some of the cases where the criterion is wrong the tests may be right. This answer to the objection must be taken only as a suggestion, for it is based upon an appeal from the ultimate reliability of the criterion, and reliability to a high degree in the criterion is a necessary presupposition for its use. Secondly, the tests may be given to the individuals, and the homogeneous classes may be formed on the basis of test performance long before either of the other criteria is available. If marks obtained in college are used, surely the average of not less than a year would have to be taken; marks from preparatory schools cannot be used because they represent evaluations of achievement based upon different standards; marks on entrance examinations are subject to errors which will be discussed later. If estimations of ability are used as criteria, the judge should have the individual judged under observation for a period of from one term to a year. Many of the advantages of homogeneous groups come in the earliest portions of college work, in exactly the periods where neither judgments nor grades can be obtained. Thus the value of mental tests will consist not so much in giving information that could not be obtained otherwise, but in giving the information *immediately*, so that the selection of groups may be made when selection is of the greatest value.

A second objection to the use of either instructors' judgments of ability or students' achievements is that these measures give no indication of the

individual's *general ability*. A business man might judge an individual in a very different manner from an instructor, and after all, which is the better able to define general intelligence? The student who secures the highest grades in classroom work may have very little ability in handling the affairs of practical everyday life. Is it possible to say that tests which correlate with criteria of the academic kind are for that reason measures of ability?

The person who makes an objection of this kind will not be content with the statement that students who secure good grades may not be interested in practical affairs; nor will the observation that the teacher's business is pretty largely concerned with the reactions of the mind, convince him that the teacher is a good judge of intelligence. And so the objection will not be met; but we shall so specify the ability under discussion that this critic may not take offense. If our tests agree with the academic criteria of ability, we shall claim only that tests may be used to select groups of homogeneous *academic* ability. This is sufficient for our purpose, since the use of mental tests in the academic situation is the object of our study. Now although instructors may not be capable of judging individuals in a way that would satisfy all standards of intelligence, they are certainly the most acceptable source of information as to the abilities of the students in academic work. If the reader is inclined to accept the academic criterion as the basis of approximating intelligence in general, so much the better; if mental tests are found to be accurate here, their applicability will be just that much wider.

One, or perhaps both, of the academic criteria may then be used to tell how the experimental groups should have been divided. There remains the examination of the relative advantages of grades and of estimations.

2

Grades, or credits, have had a wide use as measures of the ability of students. Wissler,²⁷ in early work at Columbia, selected grades as his criterion; and in the most recent reports of test work on college students at hand, that of Rowland and Lowden¹⁷ at Reed College, and of Bell¹ at the University of Texas grades are still favored. Grades, aside from being records of the actual academic *achievement* of students, have several important advantages which probably account for their popularity as an index of ability. In the first place, they are easily obtained. It is necessary that the instructor turn into the administrative office a mark showing the performance of the student in his course, and this mark once turned in is ever afterward available for the purposes of research. This availability of grades is a very important virtue, and must be seriously con-

sidered. In the second place, grades lend themselves easily to averaging. If the grade is expressed numerically the average of the student's abilities in any number of courses may be expressed directly; if the grade is expressed by letters, the institution may have some credit scheme whereby a numerical value may be attached to each letter. Averaging makes possible a partial neutralization of accidental errors, and thus greatly increases the reliability of the measure. In the third place, grades usually indicate that there are a few individuals who are very good, and a few others who are very poor, with the majority clustered about some mid-value. This is just the kind of distribution that would be expected in the case of abilities, and it takes into account the fact that the difference between two adjacent individuals at either of the extremes of ability is greater than the difference between two individuals near the mean ability. Finally, assuming that the standard of grading does not change greatly from year to year, we may combine into a single group individuals who were in reality members of successive classes. This advantage is important since it makes possible the formation of a single group of any desired size.

In spite of these many advantages, the use of grades as a criterion of academic ability is open to serious objections. Too often instructors make of the grade an administrative device, inciting certain students to greater efforts. To bright students a lower mark may be given than is actually deserved; to poor students a higher mark may be given, just as an encouragement. In these cases the grade has a complex meaning; it is no longer a simple measure of academic achievement.

Then too, the precise ability represented by any mark is not defined, so that instructors grade by very different standards. As a result, all students must be graded by exactly the same instructors, or else the grades given by all instructors must be stated in terms of the same average and same dispersion. Otherwise, the taking of an average is not allowable, and the grade criterion instantly loses one of its greatest advantages.

Finally, grades as measures of ability are subject to both accidental and constant errors. The difference between constant er-

rors and accidental errors must be carefully noted, for it will play a very important part in discussions to follow.⁸ An accidental error is an error that is produced by the interplay of many irregular and unrelated influences, and as a result the empirical measurement may be either greater or less than the true value of the measured object. As long as errors are accidental, it is possible to obtain a close approximation to the true value by determining the point where the mean square deviation of the empirical measurements is a minimum (the arithmetic mean of the measurements) and by using this point as the true value.¹⁴ The several grades that are obtained by a student may be considered empirical measurements of academic achievement, and as such they are subject to many accidental errors. These are due to such causes as variable personal reactions of instructors, and variable interests of the student.

But grades are also affected by many *constant errors*. These are errors which tend to displace an individual's total standing, since they tend to displace each of the student's grades in the same direction. Severe economic pressure, social interests, general ill health or temporary absences, tendency to nervous instability at times of recitation or examination,—these factors operate on all grades in precisely the same way. For example, if a student gives an undue amount of time to social activities, *all* of his grades will be lowered, *and no amount of averaging will even approximate a true index of his ability*. In the physical sciences, the standard method of eliminating the effect of constant errors is to determine their magnitude, and to make the necessary corrections in the measurements. Corrections for the effect of temperature and barometric pressure are readily made, but the problem of estimating how much lower a student's grades are because of the fact that he traveled for a week with the dramatic club is insoluble.

There is considerable variability in the amount of constant error in different institutions, depending upon the attitude of the student body toward the work of the curriculum.

3

The unreliability of grades as measures of academic ability, due to their use for disciplinary purposes, their indefinite meaning, and their serious constant errors, makes their use as a criterion of ability questionable.* We shall, therefore, attempt to find a satisfactory method of obtaining estimations of ability from instructors.

The method of obtaining estimations used by earlier investigators, e.g., Dressler,⁶ Gilbert,⁹ and Kirkpatrick,¹² was to ask that the student be placed in one of the three classes, *good*, *medium*, or *poor*. This division may be satisfactory if one wishes only an indication of the differences between the averages of the three classes in some particular performance, but it is valueless in specifying how well or at what point a *division* of the entire group may be made. Such a classification is based upon the assumption that all instructors will conceive identical points dividing the undefined classes *good* and *medium*, and *medium* and *poor*. Although the judgments are easy to make by this method, for our purpose the resulting divisions are entirely too rough.

A second method and a very important one is the *order of merit* or *rank method*. This method has been used in many investigations; among the most important are Cattell's⁵ studies of American men of science, and Spearman's²¹ and Burt's⁸ studies of general intelligence. The judge is usually instructed to select from the group the individual whom he considers foremost in respect to the quality in question; and to place this individual first. From the remainder of the group the best is again chosen, and is placed next in rank. Sometimes the procedure is varied, and the poorest is selected after the best. When great accuracy is desired, and there are not too many individuals, a modification of the *order of merit method*, the *method of paired comparisons*, may be used. By this method every individual is paired with every other one, and in all cases where an individual excels the person with whom he is paired, he is given a plus mark. The

* If homogeneity in performance were the sort of homogeneity desired, grades might be used as the criterion of the value of mental tests.

individuals are finally arranged according to the number of plus marks they receive. The increase in accuracy resulting from this elaboration of the method is probably not great enough to compensate for the added time required to make the judgments. Certainly the number of subjects required for an investigation of this kind precludes the use of the method of paired comparisons.

The order of merit method does not offer many difficulties in situations where it can be used, and it has some noteworthy advantages. The method permits averaging the judgments on any one individual so that the harmful effect of accidental errors may be reduced. Furthermore, the probable difference between the positions of two individuals may be calculated. Both of these advantages of the order of merit method were shown by Cattell.⁶ In his research, the ten foremost men in each science were required to place from one hundred to three hundred of their colleagues in the order of their ability. Suppose one individual received the ranks of 10 15 16 12 14 21 15 19 12 11; an average of these ranks may be computed to give his average rank. The probable error of this individual's rank is obtained from the variability of the judgments upon him. "The difference in scientific merit between any two of the psychologists . . . is directly as the distance between them, and inversely as their probable errors. If two of them are close together on the scale, and if the probable errors are large, the difference between them is small, and conversely." If the average ranks of all are determined, an order of merit may be arranged on the basis of the average ranks.

In common with grades, the order of merit method gives an exact evaluation of the ability of each individual, and we are not compelled beforehand to accept, in each of our final classes, a definite percentage of the entire group.

An objection often urged against the order of merit method is that two individuals of average ability appear as widely separated as two individuals of extreme ability. This is a valid objection to the method as it is usually used, but there are certain devices whereby the difficulty may be removed. One is Cattell's plan for

estimating the differences between individuals from the probable errors of the judgments upon them. Another, which has been used extensively by Thorndike,²³ is that of measuring the differences between two qualities on a scale in terms of the ratio of the judgments of superiority to judgments of inferiority, on one of the qualities. When the ratio is 3, the difference is called a *unit* difference. A third method was employed by Galton⁷ in his study of hereditary genius. If it is assumed that individual abilities distribute in the form of the probability curve, we may arrange our individuals in the probability curve, and determine just what point in the scale corresponds to every person. This has the effect of decreasing the differences between mediocre individuals, and of increasing the differences between individuals at the extremes. The labor involved in making this correction is considerable, but it may be lessened by the use of tables.

In spite of its advantages, the order of merit method has one limitation which has not yet been satisfactorily overcome. It requires that every judge have sufficient information about every individual to justify an estimation. For the method makes it imperative that every student be ranked by every instructor. There may be no partial lists; and the fact that a judge is not acquainted with one man is enough to discard technically the rest of his rankings. In the concrete situation, where the instructor has only 20 or 25 students in a class, his contribution to a final order of merit would be quite valueless. In researches where moderately large groups have been tested, and this difficulty has become acute, it has been met by using the rankings of only one judge, and by neglecting the records of whatever students he failed to estimate. This device loses for the order of merit method many of its advantages, and makes the ranks actually obtained subject to many errors. Unless the limitation here mentioned can be removed, the order of merit method cannot be used in mental test investigations which use judgments as their criterion and attempt to consider even a medium number of subjects.

Thorndike²⁴ has proposed a method for approximating the true order of merit where the series of judgments are partial. It is impossible to give an exposition of the method without the

reproduction of many tables, and so the reader is referred to the original article. The method, unfortunately, is not applicable to our conditions, for it requires that each individual be estimated by many judges, and during a semester it is unlikely that a student will have more than four or five instructors. It is further necessary that the different instructors judge many individuals in common, and in a freshman class that numbers more than one hundred, the required overlapping may not be found. Thorndike's method is probably applicable in certain circumstances, but it does not make the order of merit method available for securing estimations of ability of the members of a freshman class.

A further disadvantage of the order of merit method is that it does not permit the combination of two separate groups into a single group of large size.

The failure of the order of merit method is due to the fact that the position of each individual is conditioned by the qualities of the other members of the same group. It seems clear that the only way to obtain the judgments is through the use of some scale that is external to and independent of any particular group. To meet this demand there are two kinds of scales: one such as that used by Webb,²⁶ and another as used by Pearson.¹⁵

Webb's plan for forming scales is not limited to scales of intelligence, but is applicable to practically any situation where estimations of qualities, e.g., honesty, sense of humor, is desired. The following excerpt from Webb's monograph describes the method.

"The following instructions were issued to all judges:

1. Personal qualities are named and briefly annotated in this schedule. If you have any doubt as to the meaning of any of them, please ask me.

2. In the columns under each subject's name place one of the marks

+3 +2 +1 0 -1 -2 -3

for each of the qualities specified.

To avoid errors, please put the + signs as well as the —.

3. The mark +3 is for those showing a very high degree of the quality as compared with the average.

+2 is for those showing a degree of the quality distinctly above the average.

+1 is for those showing a degree of the quality slightly above the average.

0 is for those possessing the average degree of the quality for the group you are judging.

—1 is for those slightly below the average.

—2 is for those distinctly below the average.

—3 is for those showing the lowest degree of the quality as compared with the average.

4. As far as it is possible in your group of 20 men, the number of subjects receiving the above marks should be 1, 2, 4, 6, 4, 2, 1 respectively."

The limitation in the fourth point is an attempt to force all the estimations into the form of the probability curve.

Judgments made on this plan may be averaged, and the seven classes seem to offer a sufficient number of points on the scale. There are however three objections to the scale which cause us to reject it for the purposes of this research.

1. It is not legitimate to force a group of twenty or thirty individuals into any specific form of distribution. Assuming that the total population distributes according to the probability curve, we might still expect a great variety of distribution in samples from the total population which contain only 20 cases. This expectation is revealed by the high probable errors of the frequency constants of distributions containing so few individuals. Since the estimations from any one judge should not greatly exceed 20, the provision contained in the fourth point is not well advised.

2. The scale is not sufficiently objective, since the 0 mark is given to the individuals possessing an "average degree of the quality for the group you are judging." This limitation on the objectivity of the scale would not be serious if all judges estimated precisely the same group of individuals. But in our situation, it is possible that an individual might receive very different marks, not because of any difference of opinion as to his academic ability, but simply because in one case he happened to be

a member of a group of high average ability, and in the second case a member of a group of low average ability.

3. There is an assumption that the differences between the steps of the scale are equal. This may conceivably be true, but the point must be demonstrated. If the steps are not equal, many errors in the position of individuals would creep in, due to the inaccuracy of the averages of the judgments. Webb says, "The most reasonable bases appear to be given by taking the seven classes as equidistant from one another. This has the effect of making the distribution approximately normal." If Webb means that equidistant steps will make *any* distribution approximately normal, he is controverted by many facts; if he means that in his particular research equidistant steps make *his* distribution normal, we need not be too greatly impressed, for he has loaded his dice by asking that the judgments be given *according to the frequency required by the normal curve for equal steps*. Of course, under such circumstances, equidistant steps have the effect of making the distribution approximately normal. The form of the distribution of the judgments is not evidence for equal steps, but is a consequence of assuming them.

For these reasons, the Webb scale is abandoned. There still remains as a possible method of securing estimations of ability the Pearson scale of intelligence.

There are three important related facts that enter into the discussion of any frequency distribution, and that must be considered in the formation of any scale. These are 1. the form (or equation) of the distribution curve; 2. the relative distance between the steps of the scale by which measurement is made; 3. the relative frequency, or the number of individuals, at each point of the scale. For example, Galton⁷ assumed that abilities followed the probability curve, and that the steps between certain classes on a scale of abilities were equal; from these assumptions he could tell what the number of individuals in each class should be. In mental measurement, it is a familiar practice to assume that the steps of a scale of measurement are equal (a questionable assumption), to determine the frequency with which observations fall at any point of the scale, and then to draw conclusions concerning the form of the distribution.

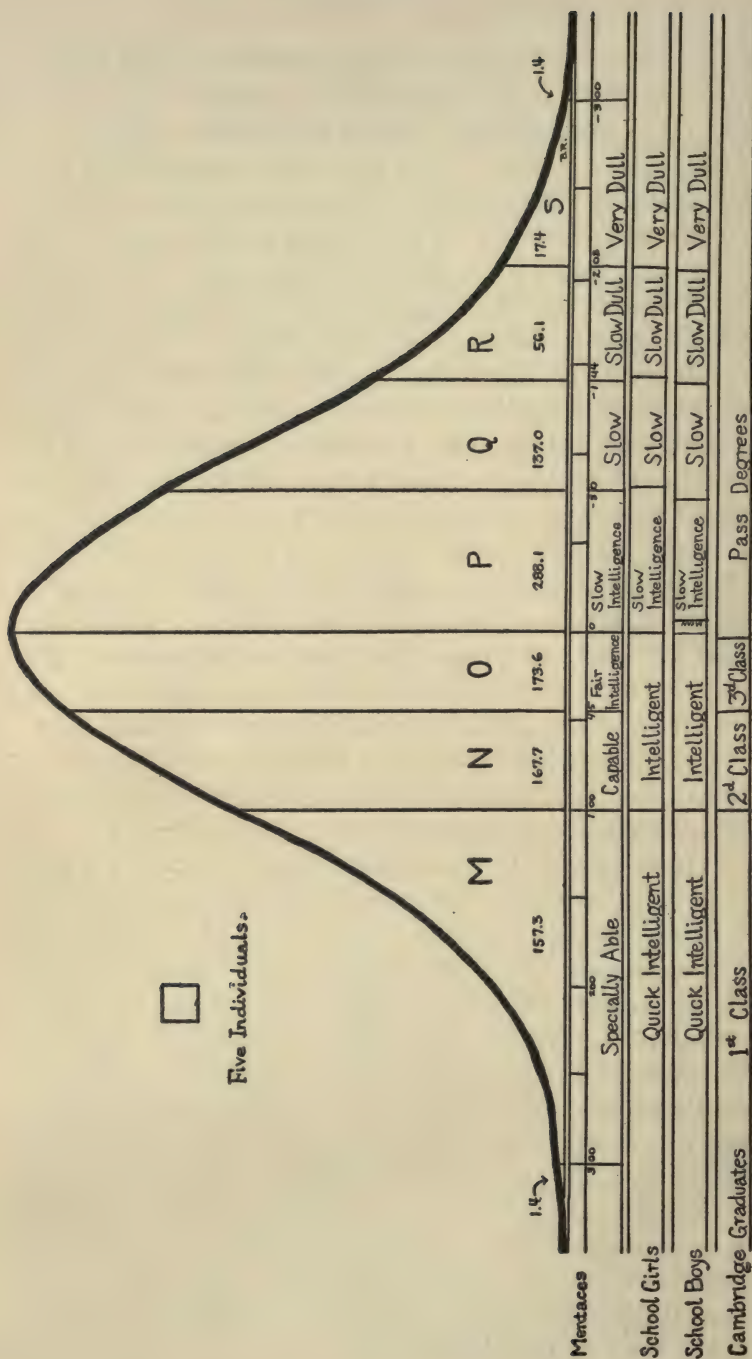


FIG. 1.—The Pearson Scale of General Intelligence. The distribution curve. For convenience, 300 mentaces is usually taken as the mean instead of 0. The standard deviation is 100 mentaces. After Karl Pearson, *Biometrika* V, 1906, p. 110.

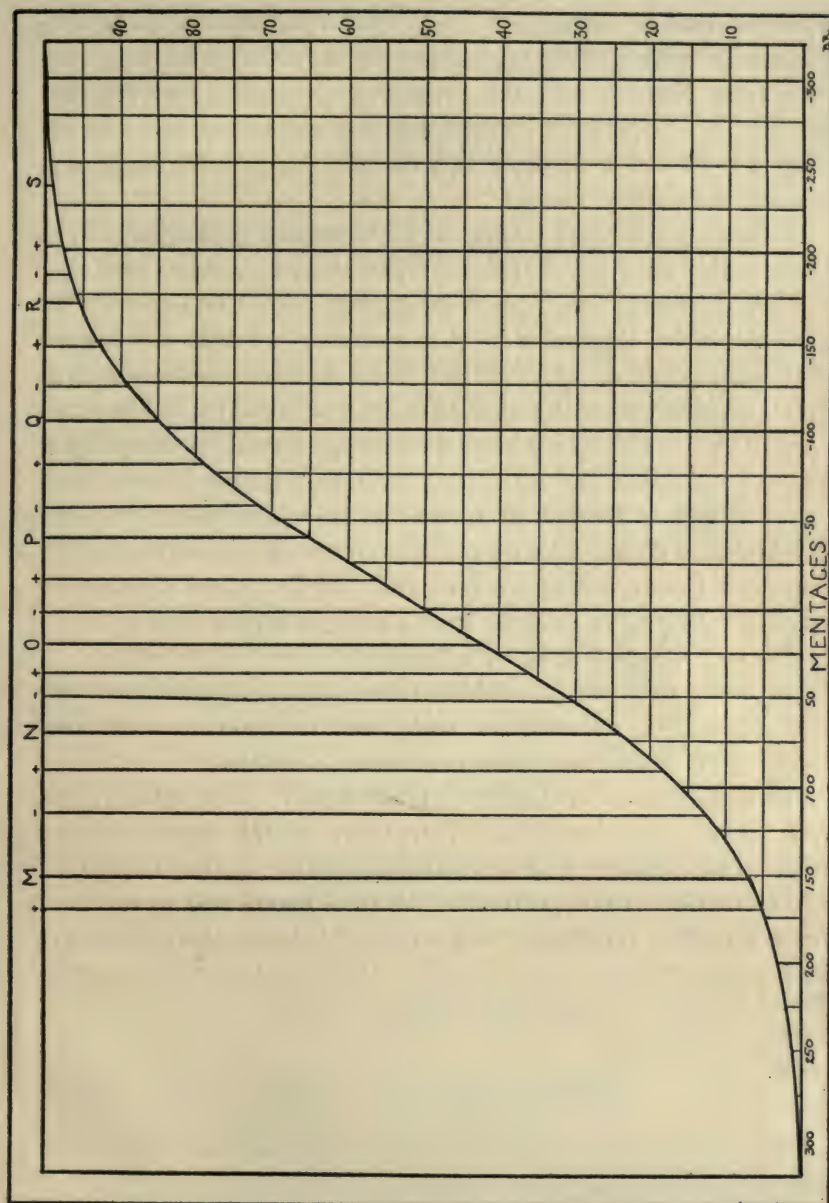


FIG. II.—Ogive of Pearson Scale of General Intelligence.

The Pearson scale was formed from the third of the possible relations: that is, intelligence in general was assumed to be distributed according to the probability curve; classes of intelligence were then defined; and the frequency of each class was determined by experiment. From these data it was possible to determine the relative distance between the steps of the scale, i.e., between the defined classes.

The assumption that intelligence follows the probability curve will probably need less justification than it did a dozen years ago. However, it seems desirable to quote Pearson on this point. "Is this assumption legitimate? It is certainly not true for organs and characters in *all* types of life. But it really does describe in a remarkable manner the distribution of most characters in mankind. We have shown within the limits of random sampling it is true for a great variety of measurements on the human skull . . . I should be the last to assert that no human characters can be found that do not diverge sensibly from the Gaussian distribution. But I believe they are few, and that for practical purposes we may with nearly absolute safety assume it as a first approximation to the actual state of affairs."

The next step was to define classes of intelligence that fall naturally into a quantitative scale, and to determine the frequency with which each class is actually estimated.

Seven classes were defined on the scale: *Very Dull*, *Slow Dull*, *Slow*, *Slow Intelligent*, *Intelligent*, *Quick Intelligent* and *Erratic-Inaccurate*. The *Erratic-Inaccurates* turned out to be only a fraction of one percent of the total group and so the class was not further considered. It was found desirable to divide the *Intelligent* class again into the *Fair Intelligent* and the *Capable*. The definitions of the classes are as follows:

Class L *Genius*.

Class M *Quick Intelligent* (Specially Able): A mind especially bright and quick both in perception and in reasoning about not only customary but novel facts. Able and accustomed to reason rightly about things on pure self-initiative.

Class N *Capable*: A mind less likely than M to originate inquiry, but quick in perception and reasoning about the perceived.

Class O *Fair Intelligent*: A mind ready to grasp and capable of perceiving facts in most fields. Capable of good reasoning with moderate effort. This group comprises, say, one third of the total population.

Class P *Slow Intelligent*: A mind slow generally, although possibly more rapid in certain fields, but quite sure of knowledge once acquired.

Class Q *Slow*: A mind advancing in general, but very slowly. With time and considerable effort not incapable of progress. Very slow in thought generally, but with time understanding is reached.

Class R *Slow Dull*: A mind capable of perceiving relationships between facts in some few fields with long and continuous effort, but not generally or without external aid.

Class S *Very Dull*: A mind capable of holding only the simplest facts, and incapable of reasoning about or grasping the relationship between facts. This group passes into the mentally defective.

Class T *Imbecile*.

Estimations were obtained by Pearson on 4638 school children of various ages. Great care was taken to gather the data from all portions of the United Kingdom, so that the sampling might be as free from local influence as possible. The data on the boys were kept separate from those on the girls so that two distinct groups were obtained.

A further group was used for reference which consisted of 1011 Cambridge graduates. These were classed in four grades: *First Class Honours*, *Second Class Honours*, *Third Class Honours*, and *Pass Degrees*.

The median point in the judgments on the boys was found to lie approximately between the *Intelligent* and *Slow Intelligent* groups. In the girls, this point was closely between the same two groups, and in the class of the Cambridge graduates it lay between the *Third Class Honours* and the *Pass Degrees*. It was further found that the frequency of the *Intelligent* group was practically equivalent to that of the *Second and Third Class Honours*; and on this basis Pearson suggests that the *Intelligent* class be divided into the two subgroups.

Since there is no reason to suppose that men and women are equally variable in intelligence, the distribution of the three groups in the probability curve was made on the assumption that the common unit of the scale for the three groups was the *Intelligent* class. This class was taken to be a common unit on the scale for the boys and for the girls; and it was equal to the range of the Cambridge graduates who received Second Class and Third Class Honours.*

It was found that when these three groups of individuals,—the boys, the girls, and the Cambridge graduates,—were distributed on the normal curve, according to the plan mentioned above, the agreement of the limits of the defined classes was very close. The three groups were combined into one, and the limits of the classes on the scale were found by determining the deviations from the average, measured in terms of the standard deviation, which were required to include in each class the observed percentage frequency. The numerical value for each class was found by determining the point on the scale corresponding to the mean value of the class.

The final units of the scale were made by dividing the range of the *Intelligent* class into one hundred parts, which are called *mentaces*. Average intelligence was put at 300 mentaces. It was then possible to express the numerical value and the range of each class in mentaces, since the relative ranges of the classes were already known. The standard deviation was found to be 93.3 mentaces. Pearson suggests, however, that 100 mentaces be taken as the standard deviation of intelligence in the formation of his scale, since this round number is a sufficiently close approximation to the true value, and lends itself more easily to calculation.

Two objections to the method of forming this scale must be mentioned. There is some reason for believing that although the form of the distribution of the Cambridge graduates may be the same as that of the children, the central tendency may be different. There must be some elimination of the lower grades of intelligence before British school children reach the university.

*15 P. 108.

This is not a serious objection, for if the scale were based on the results from the boys and girls alone, it would be practically unchanged.

It is also to be noticed that the division of the *Intelligent* class into subclasses was made after the estimations had been collected, and hence these classes as finally defined were not objectively measured. However, the ranges of the classes on either side of the *Intelligent* class were determined, and so it is probable that the only error would arise in the point of division of the two subclasses. In a later paper slight changes were made in the wording of two of the classes, but the effect of these changes upon the scale is probably inappreciable.

The question now arises whether a scale formed on the basis of estimations of children's intelligence is suitable for the purpose of securing estimations of college freshmen. If there is found no relation between intelligence as judged by the scale and age, then there can be no objection to the use of the scale. The children judged varied in age from 4 to 20, the ages between 6 and 18 each including 39 cases or more. The correlation ratio between age and intelligence was found to be for boys—.054, and for girls—.081. These relations are practically inappreciable, and the ages covered extend well into the period in which we are interested. There seems, therefore, to be no objection on these grounds to the use of the scale.

4

We have not yet examined the errors that may occur in estimations, obtained either by this plan or by any other. The errors that are really serious, the constant errors, are probably few when estimations are used, because a *variable* allowance is made when abilities are estimated for the factors that produce the constant errors in the case of grades, and errors from these sources may therefore be considered of the accidental type. If a student were very modest, and kept rigidly to a speechless part in all his work, he would probably be underestimated; and yet, a bright student of this kind would do so surprisingly well in written work that the harmful effect of his attitude might con-

ceivably be removed after a semester of study. The opposite type, the affable, diplomatic student, might be overestimated; yet the question might be raised whether a student who is able to fool all of his instructors by sheer affability and diplomacy is really so very stupid. Cases of *universal* overestimation or underestimation of a student would probably be very rare. And only in so far as the errors are in the *same* direction may they be called *constant* errors.

One important source of constant errors in the use of estimations comes through the conversations of instructors concerning students. It is probable that in a few cases, a set of judgments on an individual might be considerably in error due to a strong and expressed favorable or unfavorable reaction on the part of one of his instructors. Errors of this sort will be more frequent in a small school than in a large one; they will also be more frequent in judgments on juniors and seniors than in estimations of freshmen.

But if the constant errors are few, the accidental errors may conceivably be very, very numerous. Let us see how many sources of accidental error we can recognize. 1. Errors might arise in case different judges interpreted the definitions of the classes differently. 2. The definitions of the classes might be ignored, and the judges might evaluate the abilities of the individuals according to their own idea of intelligence. 3. A judge might feel that he knew a student well enough to form an estimate, when, as a matter of fact, the student's real ability had never been shown. 4. Judges might make too great or too little allowance for the student's social interests, his athletic ability, his nervousness in recitation, his absences. 5. There is opportunity for all manner of personal prejudice, both favorable and unfavorable. 6. A student with exceptional interest in one branch might be judged too high by one instructor and too low by every other one. 7. Besides all these sources of error, there are the countless small mistakes that are bound to occur in every situation where an estimation is required. To be sure, these are technically *accidental errors*; yet if their number and importance were too great we should have little faith in the mathematical machinery devised to cope with them.

There is an extremely simple and definite method of estimating the importance of the accidental errors. This is the observation of the size of the average of the variations of the judgments on the students. For although compensatory errors have the effect of making the difference between the average of the estimations and the true value approximately zero, *they cannot operate to make the variability of the estimations vanish. On the contrary, the variability may be increased.* A simple illustration of this is found in the following condition. Suppose an object to have a true value, 8; and suppose two measurements to be made on the object, one with the result of 8, and another with the result of 12. The average of the measurements is 10 and the average deviation or variability is 2. If a third measure which contains a compensatory error, say 4, be made, the average becomes 8, which is the true value; but the variability is increased from 2 to 3.2-3. Thus accidental errors neutralize one another in the average, but they cannot do this in the average deviation. The only accidental errors that will reduce the average deviation are those which cause a measurement very close to the average of the previous measures. If these previous measures have been influenced principally by accidental errors, their average will be near the true value, and the new measure which is near this average can consequently contain *only a small accidental error*. Thus it follows that the size of the average deviation denotes the magnitude of the accidental errors which have affected the actual estimations.

To determine how serious the accidental errors are (we grant readily that there may be many of them) we have computed the average of the variability of the judgments on an individual for two groups of college freshmen. Since some variability might be expected, due to the fact that the judges have observed the students under different conditions, it is astonishing to find that the average variability of judgments in one group containing 39 students was only three-fifths of the average difference between the major classes of the Pearson scale. In a second group of 52 students the average variability of judgments was only two-fifths

of the average difference between the classes.¹ The size of this variability may be still better judged by noting that if three judges estimated each person, and that if one judge had a tendency to estimate one class too high, *even though there were no other errors*, the variability from this cause alone would be four-ninths of the average difference between the classes. These variabilities are similar to those found by Waite who used the same scale. From Waite's²⁵ data, comprising 3427 pairs of estimations, the average variability seems to have been about three-fifths of a class interval. From this experimental work it is possible to conclude that the unreliability of the judgments due to accidental errors is probably not great, certainly not great enough to discourage the use of the scale.

During the preliminary experimental work of adapting the Pearson scale to the purposes of this research, it was at once evident that the seven classes provided by Pearson did not permit as much differentiation as the judges were able to make. Accordingly, the plan was adopted of allowing the instructor to add to the letter designating any class + or —, if the student seemed to him to belong in one class with a marked leaning in either one direction or the other. The majority of the judges were satisfied to use the straight symbols, yet the + and the — were used often enough to make their retention as part of the scale advisable. The evaluation of the + and — classes is not difficult. The mean of each major class is found, and the number of cases falling between each pair of adjacent averages is found from the probability integral. Each group of cases is then divided into three parts, and the two points of division on the scale give the numerical value of the + and — subclasses.

With the scale in this form, the following directions for making the estimations of ability were sent to instructors:

"In order to standardize judgments on mental ability, the following classification of intelligence has been selected. Please note that the scale covers the range of the population at large from the genius to the imbecile.

*Judgments on a further group indicate that two-fifths of a class interval is a lower value than will ordinarily be found. Three-fifths is more to be expected.

On the accompanying blanks, please place the letter standing for the class in which you judge the student's mental capacities to fall. What is desired is a judgment of general intelligence and not of classroom performance. If the individual seems to be in one class with a marked leaning toward another class, judge him to be in the more certain class, and indicate the direction of his leaning by + or —."

The definitions of the classes of intelligence followed. Letters M, N, O, etc., were used to indicate the various classes which could not be easily confused with the letters used for grades in the University of Chicago, i.e., A, B, C, D, and E.

No judgments were made until the conclusion of the term, in many cases of the school year. It is encouraging to note that several instructors returned blanks with the remark that they felt unable to make judgments on some students due to prejudice or lack of information.

As soon as the estimations were received the letters were translated into their numerical equivalents, and the average of the judgments and the average variation of judgments were found for each student.

The translation is made by the following table:

Table I

M+471 mentaces	N—353 mentaces	P 262 mentaces	R+157 mentaces
M 451 "	O+337 "	P—242 "	R 130 "
M—416 "	O 322 "	Q+220 "	R—116 "
N+391 "	O—302 "	Q 192 "	S+ 96 "
N 371 "	P+282 "	Q—177 "	S 62 "

The average of the judgments on an individual is the measure of his ability that is finally used.

The following is a summary of the standings of two groups. Group J_1 consists of 52 freshmen; group J_2 consists of 39 freshmen.

Table II

Groups	Mean Standing	Median Standing	Standard Deviation	Range
J_1	321.2	322	45.1	439-196
J_2	331.8	331	44.1	451-221

These results show striking likenesses. The students were

taken from two separate institutions and different standards of judging might have been expected to have influenced the average and the dispersion. The similarities are strong evidence for the objectivity of the Pearson scale.

Comparing these data with the general scale, we find that the average has been increased by about 25 mentaces. This increase is probably not so large as might have been expected. There is scarcely any tendency toward skewness, as is shown by the closeness of the median and the mean. Even the slight skewness is found to be in the opposite direction in the two cases. The standard deviation is only about half that of the population at large. A reduction in dispersion would be expected, for the extreme cases at the lower end do not occur, and the range is consequently considerably less than is that of the entire population.

The amount of objectivity which the Pearson scale seems to show might appear at first glance astonishingly great. The actual verbal definitions of the classes of intelligence seem on analysis to be susceptible to a variety of interpretations, and it might be questioned on a priori grounds whether the scale offers a method of subdividing the conceptual range of intelligence in a manner that would be uniform and definite for the different judges. The definiteness probably arises from the repeated association of the terminology of the scale with typical manifestations of intelligence, and from the rather uncritical acceptance of this terminology as an adequate identifying mark of these manifestations. The uniformity of the division of the conceptual range for different judges may come from the similarity of the environments in which the terminology acquired its meaning for the different judges. Had the judges been drawn from all walks of life, it is probable that less uniformity would have accompanied the use of the scale.

On the whole the scale gives results which seem satisfactory. We may therefore with good reason accept the Pearson scale as the best available method of obtaining judgments of the general ability of students; and we may use the judgments thus obtained as the criterion of the value of mental tests in sorting from a mixed freshman class smaller groups of greater general intellectual homogeneity.

5

It may be well to add at this point that such judgments may be put to other uses than that of testing the value of the tests. The judgments should be continued through the college course, and changes from term to term in the judgments on any student should be watched. Waite²⁵ found that in the majority of cases the change is slight, yet the exceptions will provide material for interesting and profitable study.

Single erratic judgments on students should be strictly observed. If one judge estimates an individual much higher than do the other judges, the possibility of special ability should be investigated. If one instructor estimates an individual far below the judgments of others there is an indication of an unfavorable personal reaction, and if possible, the student should be placed under another instructor.

These applications of the judgments promise that in practical work many other uses for them may be found. The judgments give information that is just as essential in properly advising the freshman as that given by the tests, and since the judgments are so easily obtained, they should form an important part of every student's record.

III

I

Although we have found a criterion of academic ability that will be satisfactory for our purpose, we must still postpone the discussion of the value of the mental tests in selecting groups of students of intellectual homogeneity.

It so happens that at present two researches must be carried on at the same time. First, we must try out many tests, and try to choose those that are the most promising. Second, we must discover how well the chosen tests *as a group of tests* will serve in the diagnosis of ability. It would be better if these two investigations could be carried on separately, but there are obvious difficulties. The evaluation of a *group* of tests is not warranted, for there is no group of tests known that would

justify an unchanged program through a period of years. The evaluation of *single* tests is impracticable, for until psychological tests are shown to have some specific administrative value, the labor involved in applying them in an institution makes their introduction for purely experimental purposes unjustified. Consequently we must point our attention in both directions at once, and if we can find indications that the tests may be used for diagnostic purposes, we may feel certain that there will be opportunity for extended research on the single tests.*

Clearly our first task is to choose groups of tests that seem most likely to give good results. This section will therefore be devoted to the selection of the groups of tests that will finally be evaluated. We do not overlook the problem of evaluating the single tests. But the methods of determining the relative weights of the single tests require that we have *accurate inter-test* correlations; and to secure these extended investigation is necessary.

So long as the tests are still in a period of development, both in respect to their content and their administration, it seems reasonable to confine research to classes of small size. Under these conditions, if a test prove worthless or a method faulty, there will be no serious loss in its discard. And yet at the same time much positive information may be obtained, even with small groups. To be sure it will be impossible to say that our results, particularly those concerning the points at which the group may be best divided, are valid for freshmen classes containing hundreds of members. But the converse is also true; for mental tests which are very satisfactory for large groups might give quite erroneous information if they were used in exactly the same way for a small and more rigidly selected group.

Two freshmen classes were tested which, though small, yielded a number of cases well above the minimum required for correlation. As will be seen from the description of the groups which follows, they are alike in little except size, and therefore, such results as are in harmony, may be taken as suggestive of what may be expected from mental tests when they are used on small freshman classes.

*22 Ch. V.

The two groups will hereafter be called J_1 and J_2 . (The letter J is chosen to distinguish these two groups for which judgments were obtained, from a group G_1 , for which grades were used as the criterion of ability.) J_1 is a first year class of the School of Commerce and Administration at the University of Chicago. The entrance requirement is high,—15 units and no conditions from an accredited high school, with an average grade of 80. Since the school gives a particular form of professional training, it is probable that the interests of the individuals are very much the same. J_1 contains 52 freshmen; of these 40 are men. The average number of judgments on an individual is 3.06, with a minimum of 2.

The second group of freshmen, J_2 , is taken from a middle western women's college. (The name of the institution is withheld by request.) J_2 contains 39 individuals. The entrance requirements for admission to this college are about the same as those for admission to the School of Commerce and Administration,—graduation from an accredited high school. Admission is still further restricted by an entrance examination. There are many differences between J_1 and J_2 . J_2 is made up entirely of women, while J_1 is for the most part composed of men. The majority of the members of J_2 have their homes in the town in which the college is located, and the group is therefore probably more homogeneous with respect to previous experience than is J_1 . The average number of estimations on the members of J_2 is 3.69, with a minimum of 2. In all, 70 students of this college were judged in order to procure information about the Pearson scale.

A third group, hereafter called G_1 , consists of Commerce and Administration freshmen. This group was tested by Dr. H. D. Kitson,¹³ and the results of this testing are the basis of his *Scientific Study of the College Student*. To him we are indebted for the use of the data. G_1 contains 50 freshmen; of these 41 are men. Unfortunately at the time this research was begun the group was so broken that it was impossible to secure estimations of the students' abilities, and so grades were used as the criterion. G_1 will be used principally for purposes of comparison with J_1 and J_2 .

There is an important difference between J_1 and G_1 that must be recognized. The attitude of J_1 toward the academic work was much more serious than was that of G_1 . The former group was given a three weeks' course in methods and ideals of study; extra-curricular activities were discouraged, and every effort was made to raise the standard of the classroom work. When we examine the relation of grades to the mental tests, this difference in attitude will be important in our interpretations.

The following tests were used in one or another of the groups.

Hard Directions	Logical Memory	Visual
Absurdities	"	" " Deferred
Sentences Built	*Constant Increment	
Opposites A	*Cancellation	
Opposites B	*Business Ingenuity	
Analogies A	*Words Built	
Analogies B	*Numbers Heard	
Alphabet	*Objects Seen	
Logical Memory Auditory	*Oral Instructions	
Logical Memory Auditory	Deferred	

The tests marked * are described in Kitson's *Scientific Study of the College Student*.^{*13} The remainder of the tests are described in Appendix I.

Table III gives the means and standard deviations of the tests for groups J_1 and J_2 . The probable errors are too high to warrant a discussion of differences between the groups.

Table IV gives the product-moment coefficient of correlation between each test and the criterion of academic ability used for the group.*

In spite of the fact that there is considerable variability throughout Table IV, the first group includes the Hard Directions, the Absurdities, the Sentences Built, the Alphabet, the Opposites, the Analogies, the Logical Memory, and the Constant

*The student's standing according to grade in courses is the average number of grade points obtained during the academic year. A grade of A gives 6 points, A— 5 points, B 4 points, B— 3 points, C 2 points, C— 1 point, D 0 points, E— 1 point, and F— 2 points. The total number of grade points of each student was divided by the number of academic units (or majors) that this student would obtain by passing all courses pursued by him during the year.

Increment tests. These tests are unquestionably the best of the nineteen for diagnostic purposes judged on the basis of their correlations with estimated ability. There is a second group of tests that appears to be extremely erratic. The tests of this group are the Cancellation and the Business Ingenuity tests, and it seems that these tests must be further investigated before they can be put on a par with the tests in the first group. In the third group are the Words Built, the Numbers Heard, the Objects Seen, and the Oral Instructions. These tests are relatively inferior.

TABLE III

Test	Group	Mean	Standard Deviation
Hard Directions Time	J ₁	115.9 sec.	31.8 sec.
	J ₂	124.8	38.5
Errors	J ₁	2.54 errors	2.08 errors
	J ₂	1.76	1.76
Absurdities ^o Time	J ₁	87.3 sec.	26.2 sec.
	J ₂	103.3 sec.	38.8
Errors	J ₁	3.42 errors	1.50 errors
	J ₂	2.84	1.13
Sentences Built	J ₂	6.7 sentcs.	2.09 sentcs.
Opposites A Time	J ₁	26.5 sec.	6.87 sec.
	J ₂	25.4	6.25
B	J ₂	43.1	13.3
Analogies A Time	J ₂	69.8	26.2
B	J ₂	69.7	24.1
Alphabet	J ₁	70.7	14.7
	J ₂	69.5	18.9
Logical Memory ^o	J ₁		
Auditory		45.5 points	17.5 points
Auditory deferred		28.6	21.8
Visual		62.5	16.1
Visual deferred		41.0	16.1
Logical Memory	J ₂		
Auditory		79.2	13.6
Visual		85.9	13.5
Constant Increment	J ₂	160.9 sec.	35.9 sec.
Number Checking	J ₁	64.3 checks	11.8 checks
	J ₂	58.2	10.1
Business Ingenuity	J ₁	28.2 points	11.7 points
Words Built	J ₁	19.6 words	3.78 words
	J ₂	20.3	4.19
Numbers Heard	J ₁	8.82 digits	1.90 digits
Objects Seen	J ₁	7.36 objects	1.10 objects

^o In this test a part of the differences in the means and standard deviations of the two groups may be attributed to differences in the test. See Appendix I.

TABLE IV

Correlation between Standing in Tests and Criteria of Academic Ability.

	G ₁		J ₁		J ₂	
	r	P.E.	r	P.E.	r	P.E.
Hard Directions Index	+ .38	.07	+ .54	.07	+ .39	.09
Absurdities Index			+ .52	.07	+ .47	.08
Sentences Built	+ .27	.09			+ .43	.09
Opposites A	+ .24	.09	+ .38	.08	+ .20	.10
Opposites B					+ .37	.09
Analogies A					+ .44	.09
Analogies B					+ .36	.09
Alphabet			+ .50	.07	+ .18	.10
Auditory Memory	+ .37	.08	+ .35	.08	+ .16	.11
Aud. Mem. Deferred	+ .31	.08	+ .45	.07		
Visual Memory	+ .20	.09	+ .26	.09	+ .47	.08
Vis. Mem. Deferred	+ .26	.09	+ .29	.09		
Constant Increment	+ .28	.09			+ .29	.10
Number Checking	+ .23	.09	+ .26	.09	— .20	.10
Business Ingenuity	+ .03	.09	+ .36	.08		
Words Built	+ .06	.09	+ .11	.09	+ .05	.11
Numbers Heard	+ .09	.09	+ .13	.09		
Objects Seen	— .08	.09	+ .09	.09		
Oral Instructions	+ .06	.09				
Standing in Tests						
Combined	+ .43	.07	+ .65	.06	+ .67	.06

No discussion of differences between the tests included in any one of these three main groups is justified because of the high probable errors of the correlation coefficients.

We are however able to choose the tests that may best be included in the groups of tests whose worth we shall finally determine.

In the series of tests which will be evaluated with respect to the estimated abilities of J₁ are included: Hard Directions, Absurdities, Alphabet, Opposites A, Logical Memory Auditory, Logical Memory Auditory Deferred, Logical Memory Visual, Logical Memory Visual Deferred. The group of tests will be called Test Series J₁.

In the series of tests which will be evaluated with respect to the estimated abilities of J₂ are included: Hard Directions, Absurdities, Sentences Built, Opposites A and B, Analogies A and B, Alphabet, Logical Memory Auditory, and Logical Memory Visual. This group of tests will be called Test Series J₂.

For the G_1 group, all the tests which were given were included in the series which was evaluated with respect to grades.

2

There are many problems concerning the actual administration of the tests that are worthy of comment. We have seen that the greatest value of mental tests comes in the *immediate* information that can be gained from them. Then, too, the rapidity with which a knowledge of the tests spreads through a college body makes it imperative that the testing be extended over the shortest possible time. These demands for speed in the application of the tests make the choice of the tests and of the conditions under which the tests are given a very important matter.

To guard against the spread of information about the tests, the psychologist should speak to the freshmen, concerning the value of accurate mental examinations. An appeal should be made to the class for its cooperation. The effectiveness of the appeal will depend partly upon the manner in which it is made, and partly upon the length of time which is required for the testing of the class. Each student should be questioned when he is tested concerning his knowledge of the tests.

The method of testing must be modified according to the size of the group that is to be tested. For classes of medium size, it is desirable to devote two periods to group tests, in order that the Deferred Logical Memory tests may be given. It is also desirable to secure measurements on the students at different times, so that disturbances from temporary indisposition may be lessened. The tests that are best given to the group as a whole are the Logical Memory tests, and the Sentences Built test. The tests that are most suitable to individual testing are the Hard Directions, the Absurdities, the Opposites, the Analogies, the Alphabet, and the Constant Increment tests. With clerical assistance, these tests may be given to an individual in about 20 minutes. The individual testing may thus be completed during the week or ten days that elapses between the first and second group tests. The data can be worked into form in three or four days; of this time, the greater part will be spent in the scoring of

the logical memory tests. It is possible in this way to secure the information from the tests within ten days or two weeks after the opening of college..

It may be that conditions will make impossible two group testings. Under such circumstances, the Deferred Logical Memory tests must be omitted. It will be possible to complete the test work in a shorter period of time if these tests are not given, and there is reason to question whether the tests for deferred memory are important enough to compensate for the delay they cause.

For very large groups consisting of from 300 to 1000 freshmen, it will be necessary to have many assistants to do the individual testing or else all the examination must be given to the class as a group. Most of the better tests are not readily adaptable for group work, and so it seems inadvisable to abandon the individual examinations. If tests are to be given in the fall, assistants may be adequately trained in a course in Mental Tests the previous spring. If assistants are employed, the group testing may be given up if it seems desirable to do so. Lack of facilities would make it necessary for the group testing to be given in parts, and there might be some difficulty in bringing enough pressure to bear upon the freshmen so that absences might be few. The student does not feel his responsibility half so keenly when examinations are made by the group plan.

The tests which might advantageously be given to large groups by individual examination are the Hard Directions, the Absurdities, the Sentences Built, the Opposites, the Analogies, the Alphabet, the Constant Increment, and the Logical Memory Visual. Since the subject does not require the attention of the experimenter in the last named test, it is possible to complete the examination in about thirty minutes. A class of 400 could be tested with the aid of ten assistants in four days; if each assistant has been trained to work up the data on the students he has tested, complete information on the entire class should be available in one week after the beginning of the fall semester. This is soon enough after the opening of college to allow for the separation of the freshmen into homogeneous classes for the work of the first semester.

The scoring of the individuals according to their performances in the tests presents as important a problem as does the actual administration of the examinations. A single numerical value for each individual's ability in all the tests combined is obtained by expressing his performance in each test as a deviation from the mean performance in the test, and by expressing the deviation in terms of the standard deviation as a unit. This gives the individual's *standing* in each test, and the sum of these *standings* is the individual's *standing* in all the tests combined. Woodworth has described this method of reducing scores to standings in detail. The labor of reducing the actual scores to standings is not excessive. The standard deviations may be easily found by the formula

$$\text{Standard Deviation} = \sqrt{\frac{\sum (\text{scores})^2}{n} - \text{mean}^2}$$

If an adding machine is used for making computations, the standard deviation may be found at the same time as the mean by printing the square of each score in a second column. The deviations may be quickly divided by the standard deviations by the use of Crelle's Calculating Tables. Of course, *if enough individuals are measured, so that grouping in a frequency table is justified, the above formula is no longer a time saving device.*

If assistants are employed in giving the tests, each assistant may report the sum of the scores and the sum of the scores squared for each test on the individuals whom he has tested himself. This information with a knowledge of the total number of students tested will make the computation of all the standard deviations the work of twenty minutes or half an hour. Tables giving the standings for each score may then be given to the assistants who can record the standings on the record cards in a very brief time.

Only individuals with complete records should be included in means and standard deviations upon which standings are based. Otherwise the great advantage of standings in computing coefficients of correlation will be lost. There will be little error in basing the standing of an individual who has not taken all the tests upon the norms of the rest of the group.

After mental tests have been used in an institution for several years, a tentative evaluation of the relative standing of the freshmen may be made almost immediately after the tests have been given, by expressing the scores in terms of the means and standard deviations of the previous years. Tables may be constructed giving the standing corresponding to any score in any test, and the clerk who records the score may record this tentative standing at the same time. The exact standing could be computed later, but it is not likely that any great variations in the relative positions of the freshmen would be found. The preliminary standings would be satisfactory for the division of the students into groups, although they could not be used easily for purposes of correlation.

IV

I

Now that we have found a method whereby the academic ability of a student may be estimated, and have selected two series of tests which, judged by the correlation between standings in a test and estimated academic ability, are relatively superior, we may proceed to the crucial question of this research: *How accurately would the tests have divided the freshman classes into groups of homogeneous academic ability had they actually been in use?*

We wish to know how a percentage—any percentage—of the individuals, representing the superior or inferior extreme of the entire class according to their performance in the series of mental tests, stands in academic ability as indicated by estimations of intelligence. More concretely, to what extent is the 15 per cent of the class which stands highest according to tests also rated highest in ability as judged? The highest 25 percent; or the lowest 15 per cent? Accuracy or inaccuracy of the internal arrangement of these groups is equally acceptable; it is only important that approximately the same individuals be chosen by the tests as are chosen by the judges.

To answer the question of this research, to determine how well

the series of tests would have done the work of separation, and for what points of division of the group the tests would have done the work most accurately, is to measure the changing relation between the group selected by the tests and judgments of academic ability, as the percentage of the total group included in the selected group is changed.

This situation presents an unusual problem in correlation. We demand a statement of the relation between a measured variable (the estimations of ability), and a second variable (the standing in the tests) which is divided at some point into two alternative categories. We should have an inadequate expression of the relation if we treated both variables as *continuous* variables, and computed an index of relationship by the product-moment correlation method, the method of rank differences, or the foot-rule. For to correlate estimations of academic ability with an exact evaluation of the performance of every individual in the test series is to measure the accuracy of the test series in terms of a problem which the test series will never be called upon to solve—namely, the determination of the precise ability of every individual. The purpose of the series of tests is fulfilled if it succeeds merely in picking out the individuals who are superior,—*it need not distinguish between superior individuals.*

2

In order to indicate the closeness of the relation in which we are interested, the relation between a *continuous* variable and a variable divided at some point into *alternative categories*, we have derived (Appendix II) a coefficient which we shall call the *rank-tangential coefficient*, and which we shall designate by the symbol t .

$$t = \frac{M(N + 1) - 2 S(R_x)}{M(N - M)}$$

where N is the total number of individuals; M the number of individuals in the selected group; and $S(R_x)$ is the sum of the ranks in variable X (the continuous variable) of the M best or worst individuals according to variable Y .

The rank-tangential coefficient proves to be a good index of relationship in several important respects:

(1) Its meaning is definite, readily understood, and adapted to the concrete situation in which t is to be used.

(2) It varies between $+1$ and -1 , taking these values only when the M best individuals in Y are associated with the M best or worst individuals in X .

(3) Its value is zero if the two variables are independent.

(4) If $M = N$, t is indeterminate as it logically should be.

(5) The rank-tangential coefficient does not measure the relation between X and Y in terms radically different from those of the common correlation coefficient. When r is low, t will be low on the average; when t is high on the average, r will be high. In chart I, values of t are plotted for different values of M . In J_2 , the coefficient of correlation between X and Y is $+.67$; in J_1 , r is $+.65$; in G_1 , r is $+.43$. In each of these three curves, the values of t are seen to vary around the value of r for that particular group.

(6) Finally, the rank-tangential coefficient is computed with great ease. Examples are given in Appendix II.

It must be emphasized that the relations measured by t and by r are different, and no direct comparisons of the two coefficients are possible, except in certain specific circumstances.

3

The formula may be applied in either of two ways; for the selected group—the variable given by alternative categories—may be taken either as standings according to tests or standings according to estimated academic ability. The coefficient at any point of division may be very different according to the way in which the formula is applied, and the interpretation of the relationship depends upon which variable, test standings or estimated abilities, forms the basis for selecting the group. If the best 25 percent in *tests* be taken, the relation found by this formula between the 25 percent group and the estimated abilities of the students tells how closely the best 25 percent *according to tests* corresponds with the continuous variable, the estimated abilities. If the 25 percent judged best in ability be taken as the selected group, the coefficient tells how closely the best 25 percent in ability corresponds to the continuous variable, now the test standings. These distinctions may seem hardly worth mentioning, but they are of the greatest importance in the interpretation of the relationships. *For if it is desired to know how well the tests would have picked out a good or a poor group, the standings in*

tests must be taken as the variable given by alternative categories. The opposite method gives information, but it does not tell how well the tests may be expected to work.

The alternative categories may be formed by dividing the group at any point according to performance in tests, and thus a series of values may be found for the changing degrees of the relationship as a greater and greater percentage of the entire class is included in the good or poor group. The values of the rank-tangential coefficient may then be plotted for 5, 10, 15, etc. percentages taken in the good or the poor group. See charts. From the values of t that are found, we may draw conclusions concerning the efficiency of the mental tests in separating good students or poor students from the total group. We may also discover at what points the group should be divided in order that the tests may do their work most efficiently.

4

Although we have a method of finding the degree of resemblance between any percentage as selected by tests and the judged abilities, we have yet to decide how close a relationship is necessary before the series of tests may be said to have a practical value. A higher or lower relation would be demanded, depending upon the inflexibility or flexibility of the system of division. For example, if it is desired to exclude students from college on the basis of tests, a much higher correlation would be demanded than if the students are to be temporarily classified for the administrator's information according to their academic abilities. As a matter of fact, the worth of a series of mental tests for the purpose of selecting homogeneous groups depends ultimately upon the degree of relation between standing in tests and ability that is thought to be necessary before a division of the class is justified.

In the report of the use of psychological tests at Reed College, Rowland and Lowden¹⁷ remark concerning a correlation of $+ .37$, "There seems to be little doubt that the revised list of tests did make a selection of the better students in Reed College." It seems probable however that a correlation of only

+ .37 would be too low to show a striking value for mental tests in selecting a homogeneous group of students.

There is an indirect way by which we may determine what degree of *correlation* would be necessary. We may arbitrarily say that the root mean square error in estimating abilities from test standings shall not exceed a given amount. We may then compute the correlation coefficient which will permit this error by the formula*

$$r = \frac{S_x^2 - S_r^2}{S_x^2} \quad \begin{array}{l} S_r = \text{root mean square error} \\ S_x = \text{standard deviation of est. abilities.} \end{array}$$

The coefficient thus obtained will be just as arbitrary as one directly chosen. But it seems reasonable to approach the choice of a critical coefficient from the point of view of the allowable error of diagnosis rather than from the more abstract mathematical relationship.

Let us take one half a class interval of the Pearson scale as a permissible root mean square error of estimation. This is about 31 mentaces. Solving the equation for r ($S_x = 43$) we obtain a value +.693. This is the product-moment coefficient of correlation that must be found between standings in tests and estimated abilities in order to justify the use of mental tests in estimating academic ability.

We shall take a rank-tangential coefficient of 70 as the minimum which would justify a division of a group at any point. This value is taken simply because 70 is suggested for the product-moment coefficient, not because there is any definite relation between the product-moment coefficient and the rank-tangential. As remarked above, *any such minimum value is highly arbitrary, depending in large measure upon the use that is to be made of the test information.*

5

The changing values of the rank-tangential coefficient were plotted in the manner described above. As a result the curves shown on charts I and II were obtained. Along the vertical axis are plotted the values of the rank-tangential coefficient; along

* Yule ⁸⁰ p. 177.

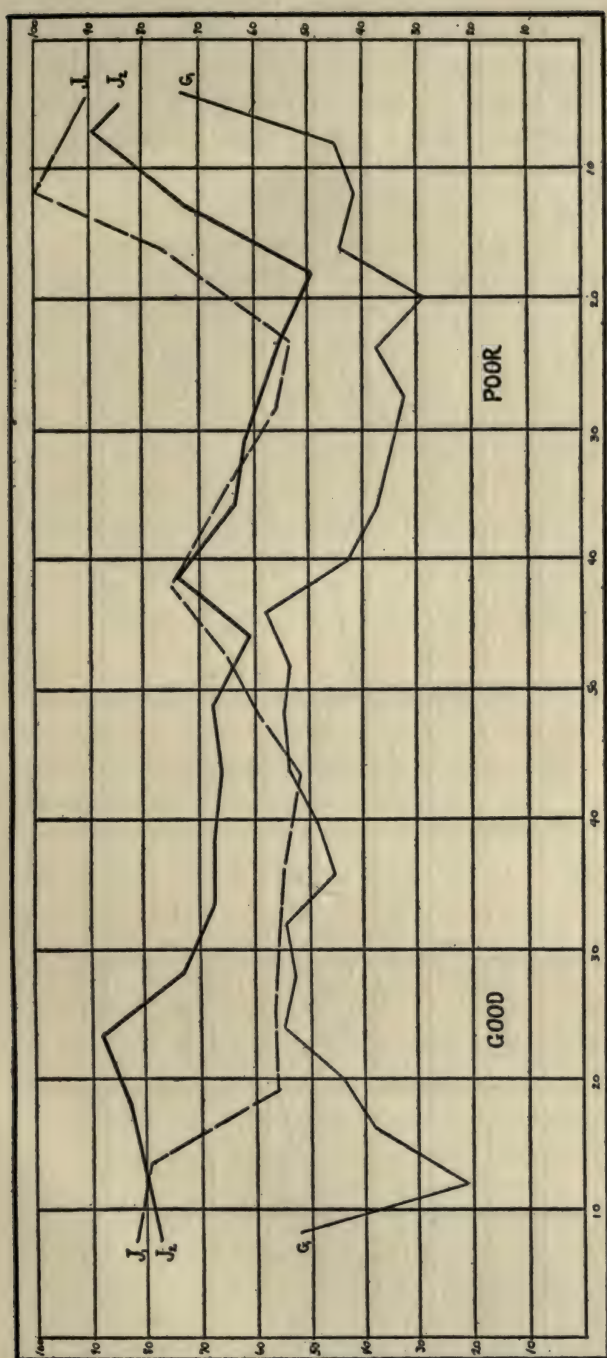


CHART I.—The relation between the rank-tangential coefficient and the percentage of individuals included in the selected group. The three curves represent three different freshman groups described in the text. The ordinates indicate the magnitude of the rank-tangential coefficient; the abscissae the percentage points of division. In each of these curves, standing in tests was taken as the variable given by alternative categories.

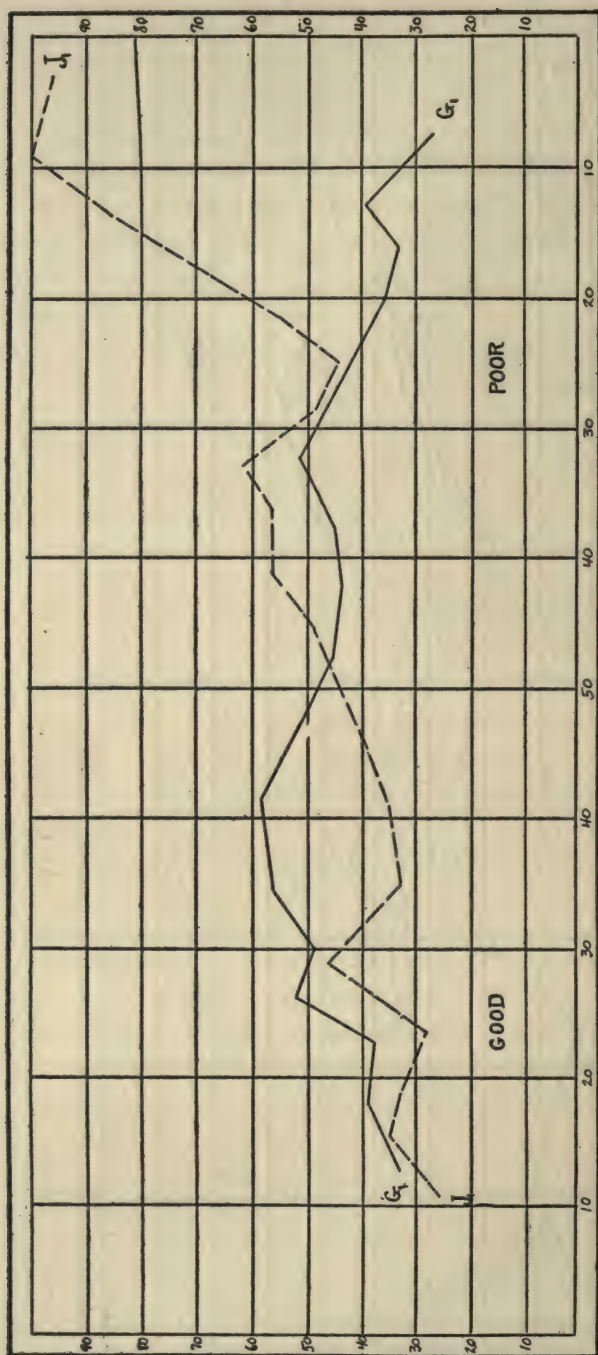


CHART II. In these curves, average grade in courses was taken as the variable given by alternative categories.

the horizontal axis are plotted the percentage points of division. The percentage is always measured from the extreme of the distribution. Thus in Chart I, 20 per cent from the left means that the best 20 per cent were included in the good group; 20 per cent from the right means that the worst 20 per cent were included in the poor group. It was found impossible to make the divisions at any regular intervals, because it often happened that two or three individuals were tied on some value, and so the division would have to be carried over the tie. Fortunately, the ties never extended over an excessive amount of the range.

Chart I shows the values of the rank-tangential coefficient for groups J_1 , J_2 and G_1 , when the group is divided according to standing in tests. Of these three curves, *the two J curves alone give trustworthy information concerning the value of mental tests in the selection of homogeneous groups*. The G curve is based on the relationship between performance in tests and grades, a criterion known to be subject to frequent and serious errors.

It must be remembered that the probable errors for t are very high (Appendix II), and hence only general similarities of the curves may be commented upon. Slight fluctuations in the value of t in one curve alone are meaningless.

In general it may be said that it is possible to distinguish four groups by the use of tests; a lowest group to include 10 to 15 per cent of the individuals; a poor group to be separated at the lower 40 per cent point, a mediocre group to include the individuals between the lower 40 and upper 15 percentage points; and the individuals who stand in the upper 15 per cent according to the tests. This is seen from the J curves, chart I.

The lower 10-15 per cent of individuals when ranked according to tests are definitely of low degree of ability. The evidence for this is especially good. Not only are the rank-tangentials for both curves high, but they reach this maximum at practically the same percentage point. It is fortunate that the lowest division includes such a small percentage of the entire freshman class. This extreme low section as tested evidently consists of those who should not be in college; it includes the majority of those

who are forced out of college at the conclusion of the first term. If we could be sure of as high a rank-tangential as that shown by J_1 , we could safely exclude the group automatically, before it begins its disastrous college career. The correlation will probably lie between the points of J_1 and J_2 , and so the group is very definitely selected. Institutions that desire to eliminate individuals of inferior academic ability would be justified in denying admission to those who are found in the lowest 10-15 per cent according to tests,—unless these individuals can present evidence of previous work of high quality.

The second low group divided at the lower 40 per cent point is clearly indicated. The rise at 43 indicates that the depression at 20-30 is due to the inclusion of slightly under-mediocre individuals in the lower section according to tests. When the percentage point at which they should have been included is reached, the curve rises. This depression and elevation indicates that the tests reverse the position of some poor and some under-mediocre individuals. There can be but few good individuals within the lower 43 per cent according to tests, for if many good individuals were included in the 20-30 depression the curve could not rise again. The fall in the curve after 43 per cent shows the presence of upper mediocres who cross the midline.

The poor group thus covers the range from 15 to 43 per cent of the lower half of the entire group according to tests. These students are the slower members of the class, and their reaction to the subject matter of the curriculum will not be the reaction of the superior individuals. This group is sharply enough defined to justify the separation of these students from the total group according to their performance in tests.

The most surprising fact shown by the charts is that the lower half of the mixed group may be divided *twice*. This fact is also the most definitely indicated, for not only are the J curves practically identical, but the grade curve of Group G_1 which is subject to many irregular influences in the lowest percentiles shows the same tendency.

The point for division in the upper half of the total group is not so clearly marked. If the J_2 curve were typical we could

make the division at any point up to 25 per cent, for the rank-tangential is well over 80. The J_1 curve is similar to the J_2 curve in shape, although it drops sooner to a considerably lower level. J_2 is probably more nearly what may be expected, for the J_2 test series is the more recent. Practically all the improvement in the J_2 series seems to have been made with reference to the good group. The depression in G_1 at 12, as will be shown later, is probably due in part to the use of grades as the criterion of ability.

The upper division may best be made at about the 15 per cent point. The height of the curves at lower 43 indicates that scarcely any of the high grade individuals can be in either of the groups, and therefore they must be in either the good or the mediocre group. The temporary misplacement of a few good students into the mediocre group should not work seriously to their disadvantage.

The lack of form of the upper judgment curves, J_1 and J_2 of chart I, as contrasted with the lower half of the curves suggests the possibility that failures of the selective agencies (entrance requirements, etc.) may have altered the character of the poorer half of the class. College freshmen as a whole are selected from the more capable of the population at large, and hence any irregularities in the selective agency will make themselves felt principally in the lower half of the freshman group. The tendency will be to make the upper section a more uniform sampling, and the elevations in lower half of chart I are quite intelligible when explained on the basis of faulty selection.

There would be little justification for dividing freshmen classes of much larger size than those here considered on the basis of the points which are indicated by the charts on the basis of this evidence. It seems likely on a priori grounds that the points at which division may be made most accurately are functions of the amount of selection which has previously operated upon the freshman group. This selection is necessarily less rigid on a group of 500 than it is on a group of 50. However, the resemblance between the two curves from groups that might be expected to be quite dissimilar, indicates that small freshman groups may be

very much the same as samples of the general population. The similarity of the curves suggests that if the points of division are carefully determined for a single large freshman class of an institution, successive classes may be divided immediately after testing, at the determined points.

6

The value of the series of mental tests in selecting homogeneous groups of students depends ultimately upon the relationship between standing in tests and academic ability; and so to find a high value for the tests, as we have, is to present direct evidence for a close relation between standing in tests and ability. But there is an indirect way of estimating the relation, namely, by observing the relation between standing in tests and the grade criterion which is known to be a faulty index of ability. If it can be shown that changes in the degree of relationship are just what would be expected if the tests actually were measures of academic ability, the evidence for tests as measures of ability is just that much stronger.

Chart II shows values of t for the relation between performances in tests and grades for groups G_1 and J_1 when the group is divided according to grades.

In general, the relation between test standing and judgments of academic ability is seen to be closer than that between test standings and grades. If tests were really a measure of intelligence, a closer relation with judgments would be expected, for on a priori grounds judgments have been shown to be a better criterion of intelligence than grades.

There are two points of general similarity in the contour of the curves of chart II. (1) In the upper half, there is a tendency to rise from a low rank-tangential to a higher one as more and more individuals are included in the upper group. This tendency stops in J_1 at 28, although in G_1 it continues to 42. This accords with the hypothesis that many high grade minds are content with resting slightly above the average grade, and therefore the highest relation with the tests of ability would not be found until these individuals are included in the group which are called good according to grades. (2) In the lower half, the curves

both show a depression in the middle of the range. The recovery in J_1 is marked; in G_1 it is so slight that it is of little significance. The depressions themselves are interesting however, for they may be due to the fact that many individuals who are capable intellectually get low grades because of social or athletic activity. This being true, a fall in the relation between grades and a measure of intelligence would be expected at the points where these individuals are found.

The differences between the two curves of chart II are partially explained by the difference in the attitude of these two groups of students toward their work. It is unfortunate that the increase of pressure that was put on the students cannot be exactly measured; the influences which were used have already been mentioned in Section III.

There are two principal differences in the curves of the two groups. (1) First and most striking is the great height to which the J curve rises in the lowest section. This is because only the students who get the lowest grades stand lowest in the tests. This is not true for the G group. If it is true that all the students in J are putting forth a great deal of work on the subjects of the curriculum, it would be expected that no students would get low grades except those who are incapable intellectually. (2) The depression in the middle half of the lower section is nearer the centre of the total group in the J curve. If one of the effects of the more serious attitude of the J group were to prevent those in J_1 who neglect their work from being content to drop so low as those of G_1 , and if the series of mental tests were actually a good method of measuring academic ability, the shift in the depression from a lower point in G_1 to a more central point in J_1 would be expected.

This indirect evidence based upon the discrepancies between the grades of the students and their standing in mental tests although not wholly unambiguous thus tends to confirm the direct evidence given from the close relation between judgments of the academic ability of the students and their standing in mental tests, that performance in mental tests is a general indication of academic ability.

V

To summarize the results of this research:

(1) Methods of estimating academic ability were examined, and the Pearson scale of intelligence was adapted so that it might be used in securing estimations of students' abilities from instructors.

(2) Two series of superior tests judged by their correlations with estimated ability were selected for evaluation with respect to their success in selecting groups of individuals, homogeneous in academic ability, from the total freshman class.

(3) A formula has been deduced to express the extent to which a certain number of individuals at either extreme of a variant Y holds that average rank position in variate X . The coefficient resulting from the use of this formula has been called the rank-tangential coefficient.

(4) The series of tests were evaluated, and it was found that they may be used with considerable accuracy in selecting groups of students of homogeneous academic ability.

(5) The points at which division may best be made were determined for freshman classes of small size, and a method was devised whereby points of division may be found for any series of tests and for groups of any size. Small groups seem best divided at the lower 10-15, the lower 40-43, and the upper 15 per cent points. The two lower points are definitely indicated; the upper point is less clearly shown, although the most recent series of tests gives a very high correlation with estimated academic ability when an upper group including the best 25 per cent according to tests, is separated from the entire class.

(6) Indirect evidence for a close relationship between test standing and academic ability was found from the changing values of the rank-tangential coefficient as the group is divided at different points.

The value of mental tests is not confined to dividing mixed classes into homogeneous groups just because they will do this work successfully. In many institutions it may be impossible or undesirable to conduct separate classes for different grades of ability. In such cases the standings in the tests may be used

simply to tell in which of the groups the student would have been placed in case the divisions had actually been made. With this information the student's advisor may talk more understandingly with the student about his difficulties, and the administrator may regulate the election of courses and the amount of extra-curricular activity of an individual with greater confidence. The more extended usefulness of mental tests in academic work should not be forgotten in the consideration of the work of separation which they do so well.

STANDING IN TESTS. J₁

150	135	120	105	90	75	60	45	30	15	0	15	30	45	60	75	90	105	120	135	150
																				460
																				445
									1											430
																				415
																				400
									1									1		385
																	2			370
															1					355
																		1		340
																	1			325
									4				2							310
									1			3								295
												1								280
																				265
																				250
																				235
																				220
																				205
																				190
																				175

Estimated Ability

STANDING IN TESTS. J₂

150	135	120	105	90	75	60	45	30	15	—	0	15	30	45	60	75	90	105	120	135	150
																		1			460
																					445
																					430
																					415
																1					400
								1													385
									1				2	1			1				370
									1				1								355
								2								1					340
										1		1	1	1							325
								2			2	1									310
								1	2	2	2	1									295
																1					280
																					265
																					250
																					235
																					220
										1											205
																					190
																					175

Estimated Ability

APPENDIX I

DESCRIPTION OF TESTS

Hard Directions. For J_1 the Woodworth Hard Directions²⁹ form without the final heavy bar was used. Because of ambiguity, the first instruction was cancelled, so that the test began, "Put a comma etc."

For J_2 a form of the test modified by Professor Woodworth was used. The changes were as follows. In line 2 the digits 2, 4, 6, 8, 9 were substituted for the letters F G H I J. In line 9 the space between *sentence:* and "*A horse* was reduced. At the end of line 22 the word *or* was added.

A form which may better be used embodies further changes which seemed necessary. In line 8, the phrase beginning *Put in a number* was changed to *Put the correct number in the next sentence*. It was necessary to introduce the idea that the number to be supplied was the right number, and that it was to be put into the incomplete sentence, not directly after the colon.

The scoring formula used was $\text{Index: } 3t + e$ where t and e are scores in time and errors respectively expressed as deviations from the mean in terms of the standard deviations as the unit deviation. Approximately this formula was given by the regression equation for both groups.

In the directions to the subject, one should emphasize that both speed and accuracy will be measured. It is also advisable to instruct the subject not to stop after the test is begun.

The test is given to subjects individually.

Absurdities. The absurdities test was constructed along the plan outlined by Simpson.²⁰ The following ten sentences are printed on strips of cardboard and are placed before the subject in a pile. Five of the sentences are logical and five are absurd.

1. Having reached the goal, I looked back, and saw my opponents still running in the distance.

2. The storm which began yesterday morning has continued without intermission for three days.

3. While sharpening his three bladed knife, my cousin cut his second finger.

4. Phyllis was born three years before her younger sister, Ruth.

5. Our office boy may get ahead at last; he was often behind before because his watch was slow, but he has been coming early of late.

6. Preferring a tarnished reputation to the possibility of becoming a corpse for the rest of his life, the young soldier took to flight.

7. Mrs. Smythe has had no children, and I understand the same was true of her mother.

8. Having dressed carefully and elaborately, she descended to the breakfast room, only to find it deserted.

9. The hands of the clock were turned back, so that the time of the sun's rising might seem later.

10. That day we came in sight of several icebergs that had been entirely melted by the warmth of the Gulf Stream.

Sentences 1, 2, 6, 8, 10 are Simpsons. Sentences 3, 5, 9, are adaptations, and sentences 4 and 7 are new. The form of the test here described is that used with J_2 with the following exceptions. In sentence 3, the word *second* has been substituted for *middle*, and sentences 7 and 9 have been interchanged.

In the sentences given to J_1 , one was too easy and another was ambiguous. The sentences have been replaced by others, and consequently the means and standard deviations of the two groups are not comparable.

In order to determine whether sentences were ambiguous in their absurdity, the subjects were asked to find their mistakes after the test proper was concluded. In no cases did subjects feel that their error was due to the possibility of more than one interpretation of a sentence.

The score is measured in time and errors. No consistent scoring formula was found for this test, although in both cases time was found to be a better measure of ability than accuracy. For J_1 an index was used, $I = 3t + e$, but for J_2 time was used alone since the formula obtained was $I = 23t + e$. It is not possible to make a recommendation for a scoring formula at this time,

but there will probably be little lost if the score is measured in time alone.

The instructions to the subject are as follows: On each of these ten cards there is a sentence. Some of these sentences are logical and others of them are absurd. An example of an absurd sentence is the following: I have three brothers, Paul, Henry and myself. (The absurdity is explained if necessary). I want you to tell me which of the sentences are sensible and which are absurd. Don't look for mistakes in spelling and in grammar, but for *logical* absurdities. If a sentence is all right, say "*Right*." If it is absurd, say "*Absurd*." As soon as you have decided about a sentence turn the card over and tell me about the next one.

It has been found that any suggestion concerning speed or a stop watch will produce an undesirable state of excitement in some subjects. If the test is given near the end of a series of speed tests, the subject will not be likely to waste any time during the progress of the test.

The test is given to subjects individually. Since the sentences are not of equal difficulty, the test is hardly adapted for group work.

Sentences Built. Three words, *citizen*, *horse*, *decree*, were given to the subjects, with the instructions to construct as many sentences as possible from these. A time limit of five minutes was set. The score in the test was the number of sentences constructed. If the last sentence was only partially complete, it was counted as a complete sentence. It is recommended that the word *petition* be substituted for the word *decree*. Occasionally a subject will not know the meaning of the latter word.

This test was given as a group test.

Opposites A. The words used in this test were

long	dead	east
soft	hot	day
white	asleep	yes
far	lost	wrong
up	wet	empty
smooth	high	top
early	dirty	

In timing the test, the watch is started with the first response. No scoring formula can be recommended. For J_1 , index = 2 $t + e$ was used; for J_2 time alone was taken.

This test is given to subjects individually. The responses are oral.

Opposite B. This opposites test consists of a series of words of slightly greater difficulty than list A. The procedure is exactly the same. The words:

good	push
north	over
heavy	young
less	city
sharp	wild
sick	rich
big	open
weak	war
come	sell
male	innocent

In view of the recent work of King & Gold¹¹ on this test, no recommendations are made concerning the content of the lists.

The score in this test was time alone.

Analogies A and B. This is the mixed relations test described by Woodworth.²⁸ Series A is the card beginning *good: bad:: long*—Series B; *eye: see:: ear*—.

The score for these tests was time of the interval between the first and last responses.

This test is given to subjects individually.

8. Alphabet. The alphabet or alphabet sorting test was used by Burt.³ The materials required are two complete alphabets of letters, each letter printed on a single card. The size of the card is one square inch. The letters may be readily obtained in a game called *Anagrams* published by Parker Brothers, Salem, Mass. The letters are then numbered from 1 to 52 on the backs of the cards, and the same order is always used. This order is:

Z P Q K T C M Z T L H R E O B L
W I K P Y B V A D N E J U R C G S V
F U X J S A W Y X G N I D Q F H M.

The letters are covered by a cardboard screen, and the subject is seated before it. The letters are not exposed until the test begins.

Each subject is given the following instructions orally: Under this screen there are two complete alphabets of letters—that is 52 letters—all mixed up. The two alphabets are just alike. Now I want you to pick up letter A, and put it here; (experimenter indicates position) then take letter B and put it beside letter A; then letter C; and so on until you have one complete alphabet. You must be sure to pick up letters in alphabetical order; you mustn't pick up letter H until letter G is placed. I want you to do this as fast as you can, and I think you can do it faster if you arrange the alphabet in two rows;—go from A to M in the first row, and from N to Z in the second,—so that you won't have to reach so far. Use both hands if you wish.

The screen is removed, and as soon as the experimenter sees that it is no longer obstructing the vision of the subject, the stopwatch is started. The watch can be carried in the right hand while the screen is being removed, and thus the time of starting may be secured with great accuracy. The watch is stopped as soon as the letter Z is placed in position.

If subjects ask questions before the experiment begins, answers are given. Errors occur rarely, and no penalty is given on this account. Subjects must be held strictly to picking up the letters in alphabetical order.

9, 10, 11, 12. Logical Memory Tests. The passages used in these tests are given in *The Scientific Study of the College Student*. For J₁, the procedure followed, and the scoring system adopted, were exactly those described by Dr. Kitson. For J₂ the tests were given as group tests, but the scoring was changed so that in the Auditory Memory test the credit given for the main sections was 5, 10, 10, 25 units respectively, with partial credit for partial reproduction. In the Visual Memory Test, the credit for the main sections was 5, 15, 15, 15. Each idea reproduced was given 1-3 unit; and an individual's total score was found by adding the units credited for ideas to the units credited for the principal sections of the passages.

Neither these passages nor this scoring formula will probably be found in the final statement of the logical memory tests. The indications are that credit should be given for the actual number of words written, instead of for the main passages. The evidence is that each word written should receive half the credit that is given for an idea correctly reproduced. However, any scoring formula that could be recommended would be practically an arbitrary one. The absence of an ultimate scoring formula should not cause the abandoning of the tests, for even with the tests in this crude form good results were secured.

For J_2 , the length of time between the first and the deferred reproduction was changed to three weeks. The results were very unsatisfactory. If the deferred logical memory tests are used, the period between tests should not be greater than two weeks.

APPENDIX II

Pearson¹⁵ has deduced a formula which offers a method of expressing the relation between a measured variable, and a second variable which is divided at some point into two classes. The meaning of the formula may be better understood from the diagram. It is assumed that the second variable would show a Gaussian distribution if measurements were made, and that the regression of X on Y is linear. When the means of the two variables are taken as 0,

$y = b_1x$, where b_1 is the regression-coefficient of X on Y. That is, $\frac{y}{\sigma_y} = r \frac{x}{\sigma_x}$.

The problem is to evaluate x and y in the diagram. x is

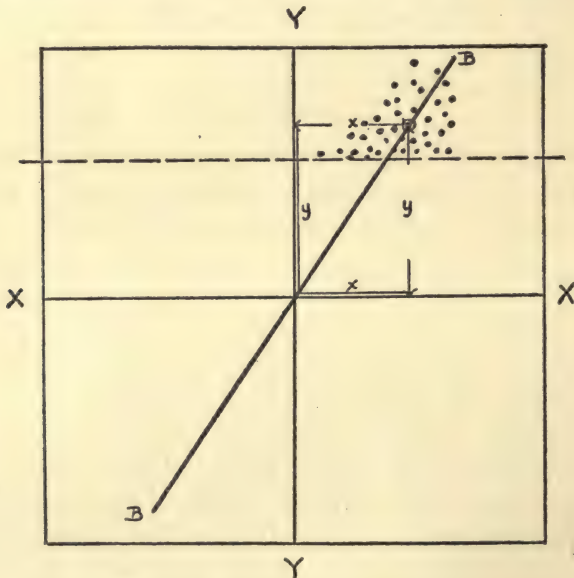


Diagram I.

easily found; it is the mean value of the measured variable which is found in one of the classes of the alternative variable. Y is

the distance from the centroid of the class of Y in which the average of the measured variable was taken, to the mean of Y. The assumption of the Gaussian frequency for the alternative variable makes possible the computation of this distance from a knowledge of the proportion of the alternative variable which is included in the class in question.

$$\frac{y}{\sigma_y} \text{ thus equals } \frac{\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y}{\sigma_y}\right)^2}}{\frac{1}{\sqrt{2\pi}} \int_{\frac{y}{\sigma}}^{\infty} e^{-\frac{1}{2}y^2} dy}$$

The steps in the computation of r by this formula are given elsewhere.¹⁸

If both variables are given in actual measurements, the limitation of Gaussian distribution in the alternative variate is removed, since for the value of $\frac{y}{\sigma_y}$ we may use the observed value for the distribution of the alternative variate with which we are dealing. However, if the distribution is closely Gaussian, the original formula may be used with considerable saving of time.

The second assumption made by Pearson is that the regression of X on Y is linear. Then "if a volume of the frequency be cut off from the frequency surface by a vertical plane at a given value of the variate Y, the vertical through the centroid of this volume cuts the regression line." If r is to have the same value for any point at which the frequency surface may be cut by the vertical plane, the vertical through the centroid of this volume must cut the regression line for any of these points of division. But this latter condition can be met only when the mean value of X associated with every Y array however small also lies on the regression line. Such strict linearity is rarely found, even though the correlation table as a whole may seem to indicate clearly a linear regression. Consequently, a certain fluctuation in the values given by the formula for different points of division of the Y variate is to be expected. *The psychological interest lies in determining whether the fluctuations are the same for differ-*

ent samplings of the same population, and in discovering what causes may lie behind them.

The actual relation between the two variables, X and Y, does not change no matter where we may divide the frequency surface, and therefore it seems confusing to call coefficients obtained by this method *coefficients of correlation*. It is conceivable that a regression might test linear according to the Blakeman² criterion, and yet hardly a single coefficient computed by this method be equal to the product-moment coefficient. For this reason it is suggested that the coefficient be called the tangential coefficient, and that it be designated by the letter T. The name tangential is selected since the coefficient is the slope of the line connecting the vertical through the centroid of the solid cut from the frequency surface, with the axis along which the Y variate is represented (when the variates are measured in terms of their standard deviations). T is equal to r when the vertical through the centroid of the solid cut off cuts the straight line which best fits the means of the X's associated with Y arrays. It is equivalent to r in meaning since it gives the mean value of X deviations which are associated with the mean of a Y array.

T is a good measure of relationship between the two variables, since high values of T indicate that the average of X variations associated with a Y array is near the mean of the Y array, and when the regression is roughly linear this can happen only through a close relation between the variables. But when the regression is markedly skewed, T may exceed + or -1 for very small selected classes, and thus fail to give any idea of relationship.

It is also apparent that T is not exactly suited to our conditions, since it depends for its value upon a ratio of *measurements* in *both* variables. Now clearly in the variable given by alternative categories, we are not interested in measurements. We assign a certain percentage of the total group to one class, and the rest of the group to the other. Thus we need only know *which* individuals are included in each of the alternative categories *regardless* of what their relative positions may be. However, if we wish to change the size of the selected group, it is necessary

to know the *order of merit* of the individuals in the variable according to which selection is to be made; for in the first division we may take the superior 10, in the second division the superior 12 etc. *Consequently, our purposes are fulfilled if we but know the rank position of each individual in the variable given by alternative categories. A completely satisfactory coefficient will take this fact into account.*

In the case of the continuous variable, the measurements themselves might be used; yet, for the sake of uniformity with the alternative variable, and in order to attain the greatest simplicity in computation, it seems desirable to use rank position in the continuous variable as well.

THE RANK-TANGENTIAL COEFFICIENT

To overcome the difficulties of the tangential coefficient and to provide an index of relationship more in harmony with the concrete situation, a formula has been deduced to express the extent to which a certain number of individuals at either extreme of the Y variate holds that average rank position in the X variate. The relation between the variables is thus measured in terms of likeness in the ranks of the individuals instead of similarity in actual measurement.

A coefficient parallel to the tangential coefficient would be given by expressing the ratio of the mean value of X deviations in rank associated with a Y array, to the mean rank deviation of that Y array. See diagram. Since the number of cases in X equals the number in Y, and since measurement is in terms of rank, the standard deviations of the variables are equal, and may be neglected.

Let the number of individuals upon which the measurement of the relation between the two variables is based be N; and let the variable Y be divided so that M individuals at one extreme of the distribution are separated from the remaining N—M.

The mean of each variable is $\frac{N + 1}{2}$; and the mean of the Y

array is $\frac{M + 1}{2}$. The line y in the diagram is therefore equal to

$\frac{N+1}{2} - \frac{M+1}{2}$; that is $\frac{N-M}{2}$. This is the *Y* deviation.

The *r*th *X* deviation is equal to $\frac{N+1}{2} - X_r$. Therefore the mean of the *X* deviations found in the *Y* array of *M* cases must be $\frac{N+1}{2} - \frac{\Sigma(R_x)}{M}$. This is the mean rank deviation for the *X*'s in the *Y* array.

The required ratio is

$$\frac{\frac{N+1}{2} - \frac{\Sigma(R_x)}{M}}{\frac{N-M}{2}} = \frac{M(N+1) - 2\Sigma(R_x)}{M(N-M)} = t$$

where $\Sigma(R_x)$ is the sum of the ranks in variable *X* of the *M* best or worst individuals according to variable *Y*.

For the sake of simplicity of treatment, the individual who makes the best performance in a test is given rank 1, regardless of whether the numerical value of a superior performance is relatively high or low.

The sign of the denominator depends upon the particular extreme, the good or the poor, of the distribution which is under consideration, and not upon the numerical values of *N* and *M*. For example, if the superior individuals of the *Y* variate are taken as *M*, *y* is a positive deviation, and the denominator is positive. If the inferior individuals of *Y* are considered, *y* is a negative deviation, and the denominator is negative.

It is suggested that *t* be called the rank-tangential coefficient. *t* proves to be a good measure of relationship. Its meaning is definite and readily understood; it varies between +1 and -1; its value is zero if the two variables are independent; if *M* = *N*, *t* is indeterminate. Finally, the coefficient is easily computed.

The following illustrations show how *t* is computed.

I. Let *N* = 10; *M* = 4. Divide according to the *Y* variate, selecting the four superior cases. Suppose the ranks of these

four cases in X are 1, 3, 4, and 6. Substitute in formula $N = 10$; $M = 4$; $S(x) = 14$

$$\frac{4(10 + 1) - 2:14}{4(10 - 4)} = +.66$$

The denominator is positive since M includes the superior cases of Y .

II. Let $N = 20$; $M = 12$. Let M include the inferior cases, and let their ranks be 6, 7, 8, 9, 11, 12, 13, 14, 16, 17, 19, and 20. $S(x) = 152$.

$$\frac{12(20 + 1) - 2:152}{12(20 - 12)} = \frac{-54}{-96} = +.542$$

The denominator is negative since M includes the 12 inferior cases of Y .

The rank-tangential coefficient is not a method of approximating the product-moment coefficient; it is a method of evaluating a relation of another sort. To the writer's knowledge the only method which has been used previously in evaluating such relationships is that of stating the percentage of identical cases, a method which is inadequate because of its ambiguity. It is only because the rank-tangential coefficient does give an index of this type of relationship that the writer feels justified in introducing it into a literature already crowded with coefficients of one sort or another.

The rank-tangential coefficient has properties similar to the product-moment coefficient of correlation:

(1) The rank-tangential coefficient gives the mean of X deviations associated with the mean of Y arrays. In this special case, however, there are only two arrays.

(2) The rank-tangential coefficient varies from $+1$ for a complete relation between the two variables, through 0 for independence to -1 for complete dissimilarity.

The writer is not able to state the probable error of the rank-tangential coefficient. Judging from the probable error of the bi-serial correlation coefficient (which is equivalent to the tangential coefficient in the case of a normal correlation surface) the probable error of the rank-tangential coefficient will be from

20 to 200 per cent larger than the probable error of the product-moment coefficient. The probable error of the bi-serial correlation coefficient increases very rapidly as one group decreases so that it includes less than 10 per cent of the total frequency.³¹

REFERENCES

- ¹ Bell, J. C. Mental Tests and College Freshmen. *J. Educational Psychology*. 1916. VII, p. 381.
- ² Blakeman, J. On Tests for Linearity of Regression. *Biometrika*. 1905. IV, p. 332.
- ³ Burt, C. Experimental Tests of General Intelligence. *British Journal of Psychology*. 1909. III, p. 94.
- ⁴ Carey, N. Factors in the Mental Processes of School Children. *British J. of Psychology*. 1916. VIII, p. 170.
- ⁵ Cattell, J. M. American Men of Science.
- ⁶ Dressler, F. B. Psychology of Touch. *American Journal of Psychology*. 1895. VI, p. 343.
- ⁷ Galton, F. Hereditary Genius.
- ⁸ Gauss, C. F. *Methods des Moindres Carres*. Trans. par J. Bertrand. 1855.
- ⁹ Gilbert. Mental and Physical Development of School Children. *Yale Studies*. 1894. II, p. 40.
- ¹⁰ Hollingworth, H. L. Vocational Psychology. 1916.
- ¹¹ King, I. and Gold, H. A Tentative Standardization of Certain Opposites Tests. *J. Educational Psychology*. 1916. VII, p. 459.
- ¹² Kirkpatrick, E. A. Individual Tests of School Children. *Psychological Review*. 1900. VII, p. 274.
- ¹³ Kitson, H. D. Scientific Study of the College Student. *Psychological Monographs*. XXIII, No. 1. 1917. Whole No. 98.
- ¹⁴ Merriman, M. Theory of Least Squares. 1894.
- ¹⁵ Pearson, K. Relationship of Intelligence to Size and Shape of Head. *Biometrika*. 1906. V, p. 105.
- ¹⁶ Pearson, K. A New Method of Determining Correlation. *Biometrika*. 1909. VII, p. 96.
- ¹⁷ Rowland, E. and Lowden, G. Psychological Tests at Reed College. *J. Experimental Psychology*. 1916. I, p. 211.
- ¹⁸ Ruml, B. Measurement of the Efficiency of Mental Tests. *Psychological Review*. 1916. XXIII, p. 501.
- ¹⁹ Ruml, B. On the Computation of the Standard Deviation. *Psychological Bulletin*. 1916. XIII, p. 444.
- ²⁰ Simpson, B. R. Correlation of Abilities. 1912.
- ²¹ Spearman, C. General Intelligence. *American J. of Psychology*. 1904. XV, p. 201.
- ²² Stern, W. *Differentielle Psychologie*. 1911.
- ²³ Thorndike, E. L. Mental and Social Measurements. 1913.
- ²⁴ Thorndike, E. L. Combining Incomplete Judgments of Relative Position. *J. Phil. Psych. and Scientific Method*. 1916. XIII, p. 197.

²⁵ Waite, H. Estimations of the General Intelligence of School Children. *Biometrika*. 1911. VIII, p. 79.

²⁶ Webb, E. Character and Intelligence. *Brit. J. of Psych. Monograph Supplements*. 1915. I, No. 3.

²⁷ Wissler, C. Correlation of Mental and Physical Tests. *Psychological Monographs I*. 1901. No. 3.

²⁸ Woodworth, R. S. Combining the Results of Several Tests. *Psychological Review*. 1912. XIX, p. 97.

²⁹ Woodworth, R. S. and Wells, F. L. Association Tests. *Psychological Monographs*. 1911. XIII, No. 5.

³⁰ Yule, G. U. Introduction to the Theory of Statistics.

³¹ Soper, H. E. Probable Error of the Bi-serial Correlation Coefficient. *Biometrika*. 1915. X, p. 384.

Vol. XXIV
No. 5

PSYCHOLOGICAL REVIEW PUBLICATIONS

Whole No. 106

THE Psychological Monographs

EDITED BY

JAMES ROWLAND ANGELL, UNIVERSITY OF CHICAGO
HOWARD C. WARREN, PRINCETON UNIVERSITY (*Review*)
JOHN B. WATSON, JOHNS HOPKINS UNIVERSITY (*J. of Exp. Psychol.*)
SHEPHERD I. FRANZ, GOVT. HOSP. FOR INSANE (*Bulletin*) and
MADISON BENTLEY, UNIVERSITY OF ILLINOIS (*Index*)

STUDIES FROM THE PSYCHOLOGICAL LABORA-
TORY OF THE UNIVERSITY OF CHICAGO

The Analysis of Mental Functions

BY

CURT ROSENOW, PH.D.

PSYCHOLOGICAL REVIEW COMPANY

PRINCETON, N. J.

AND LANCASTER, PA.

AGENTS: G. E. STECHERT & CO., LONDON (2 Star Yard, Carey St., W. C.)
LEIPZIG (Koenig Str., 37); PARIS (16 Rue de Condé)



[11]

A graceful custom offers me the welcome opportunity to give expression to a part of the sense of indebtedness which I feel. If this little work be found to have any merit at all, I believe that merit to lie largely in the point of view from which it is written. There is a sense in which that point of view is my own. There is another sense in which it is entirely due to the minds with which I have had contact. In that sense, it is my privilege to express my obligation to William James, whom I was not fortunate enough to know in person, to James R. Angell, to George Herbert Mead, and to Addison W. Moore. Many others have given their grateful testimony with regard to the spirit of scientific comradeship which prevails in the department over which Professor Angell presides. I wish to add to this my appreciation of his rare gift of appreciating and encouraging many divers and divergent types of mind. I feel that he who cannot work out his intellectual and psychological salvation here, can do so nowhere.

My thanks are due to Prof. Harvey A. Carr for thorough and stimulating instruction. I wish to say, furthermore, that my interest in the subject of modern statistical method is due to my contact with the enthusiasm and the ability of Dr. Beardsley Ruml. And I wish to voice my gratitude to Dr. H. D. Kitson whose kindness and scientific courtesy in furnishing me with data has made this little study possible.

CURT ROSENOW.

University of Chicago.

June 8, 1917

I. THE PROBLEM.

Whipple, in the introduction to his Manual, tells us that a "mental test" is the experimental determination for a given individual of some phase of his mental capacity, the scientific measurement of some one of his mental traits. He goes on to say that its purpose is practical and diagnostic rather than theoretical and analytic, but he recognizes that theory and analysis are likely to interact with practice and diagnosis in a way which is beneficial to both. Then, with a candor which cannot be praised too highly, he tells us that there is as yet no such thing as a science of mental tests. ". . . there is, at the present time, scarcely a single mental test that can be applied unequivocally as a psychical measuring rod. . . . we too often do not know what we are measuring; and we too seldom realize the astounding complexity, variety and delicacy of form of our psychical nature." And finally we are told that the pressing need of the day is not the inventing of new tests, but the exhaustive investigation of those we already have.

I cannot subscribe too heartily to some of these sentiments. To be sure, Whipple thinks that rigid standardization and the setting up of norms is the most urgent need, whereas I believe that new tests of the right kind and evaluation of old tests, is what the situation calls for. It is all too true that we do not know what we are measuring, and it seems a rather futile pastime to standardize the measures of we know not what. We need to devise tests so that we *do* know what we are measuring, and in order to do so we must subject our so-called tests to intensive study and analysis. Only so can standardization proceed intelligently.

What is it that we *are* measuring? In the first place, and it seems to me that some of our mental testers have quite lost sight of this obvious fact, we are measuring actual, factual performance at some definite specific task, like thinking of the "opposites" of a list of adjectives, or recalling the names of a number of

familiar objects which have been seen during a brief interval of time. In the second place we postulate, assume, or fondly hope that this actual performance is symptomatic of ability in some wider range of mental ability which we call a mental function.¹ We hope that excellence at the "opposites" test is symptomatic of ability quickly to recall the appropriate associate in a much wider range of situations than the test situation. We expect that ability in the "Objects Seen" test is indicative of the capacity to perceive quickly and accurately a large number of perceptual factors any one of which may later become significant. At any rate we entertain some such hopes as this unless we are willing, stupidly, to *define* "associative power" as ability in the "opposites test," and "quickness and accuracy of perception" as ability in the "objects Seen" test. But how are we to know whether our postulates are justifiable, our hopes fulfilled? Where is "Speed of Association," "Quickness of Perception" to be found?

The answer, it would seem, is correlation. We must correlate performance in tests with performance in some larger line of activity which clearly is indicative of ability of one kind or another. It might perhaps be conceivable that we should find activities, suitable for this purpose, which are more or less clearly symptomatic of the various categories of the psychologist or even of the mental tester. But be that as it may, practically it is quite impossible. We are face to face here with another of the difficulties to which Whipple calls attention, the infinite complexity and variety of life. We cannot deal with function directly. We are obliged to deal with concrete specific fact. Furthermore the need of a mass of quantitative data confines us to the fields in which men engage 'en masse' and, preferably, where quantitative data are available. Practically, the science of Mental Tests is in its infancy in the field of education, and is in embryo in psychiatry, criminology, and industry.

The data for this study were gathered in the field of education and we may proceed at once to the consideration of the criterion with which we are to correlate. We consider only two criteria.

¹ The term "function" is used loosely to indicate almost any kind of psychophysical process.

They are *academic marks*, and *judgments of "intelligence."* The latter are usually given by the same instructors who give the marks and occasionally by other judges alleged to be competent. Marks are a perfectly straightforward, objective form of achievement and, in spite of the many objections which can be made to them, they have been used extensively on account of their availability and on account of the quantitative form in which they are immediately given. Judgments have been used chiefly by those investigators who are dissatisfied with marks as a measure of intelligence, and they are supposed to reach "intelligence" in more direct if less objective fashion. Even the objection of lack of objectivity tends to disappear when the scale according to which judgments are made is the one originated by Karl Pearson to which Mr. Ruml has called the attention of psychologists in most interesting fashion.² However the choice of a criterion depends on the duty we expect that criterion perform, and inasmuch as my ideas on this subject are very different from those of Mr. Ruml, even though I also am inclined to prefer judgments to marks, it will pay us to look into that phase of the matter.

Mr. Ruml argues that the criterion is, so to speak, an absolute criterion. If it does not measure "intelligence," it at any rate defines it, and thus becomes the sole measure of the value of the tests. If we do not take this position, he would urge, we are reasoning in a circle. If we decide on a given criterion because it correlates more highly with a given set of tests than some other criterion, we are choosing the criterion on the basis of the tests rather than the reverse. Now such considerations are perfectly sound if one is interested in *measuring* intelligence. The present writer is not. He prefers to analyze it, and he conceives the analysis of the factors which enter into some objective performance such as the obtaining of marks or judgments from an instructor to be so useful and interesting a prolegomenon that perhaps it may render the more ambitious task superfluous. For this reason he prefers the criterion which gives the highest correlation with a given set of tests, for high correlations give the analyst something to work with, whereas, as we shall see, it is next to impos-

² Psychological Monograph, No. 105.

sible to reason from low correlations. The writer believes that judgments will be productive of higher correlations with "good" tests (note the circle) because he thinks that in practice judgments amount to little but a revision of the marks and the correction of a few very flagrant cases, where achievement most evidently differs from ability to achieve. However this is a matter for future research to decide. Whatever criterion gives the highest correlation with tests will be best for purposes of analysis quite apart from *a priori* considerations. In the present study marks were used as a criterion because judgments were not available. Otherwise a comparison would have been made.

If now we review the ground we have been over, it would seem that we are in a difficult position. On the one hand we have urged that tests should not be credited with symptomatic value for *a priori* reasons. On the other hand we have alleged that it is practically impossible to find other facts which do possess such value. How then, we may be asked, can one set of facts whose meaning is unknown to us serve as a measure for another set about the significance of which we are equally ignorant? If the conditions have been represented correctly, the problem is indeed difficult. And yet it is precisely this problem with which the present paper deals. So it will be worth while to dwell a little on the question whether the problem does indeed take this form, even though, to many minds, the position will scarcely need defence.

To anyone at all acquainted with the canons of scientific evidence it will be obvious that the good intention of the psychologist who devises a "test" cannot be accepted in lieu of its symptomatic value. We may grant that the psychologist has knowledge of the workings of the human mind which, in some respects, goes beyond that of the man of affairs. We cannot grant that this knowledge, basic and fundamental though it may be, makes him a competent judge of specific social and practical efficiency or of general intelligence. To be sure, perception, memory, imagination, association, attention, generalization, etc., enter *formally* into every operation of the mind. But the relation of *content* to *form* is one of the major *problems* of education and of

psychology, not one of its fundamental laws whose formula we know and can apply with confidence. The whole problem of transfer of training and of formal discipline confronts us here. *Some* transfer, or better some identity of function in divers activities, is indeed a presupposition which is necessary so that tests may be thinkable. But to ascertain the kind and degree of identity constitutes a problem which can be solved only by investigation.

Nor are things otherwise when we come to the consideration of the criterion. Whence can the authority be derived by virtue of which we credit any set of facts whatever with being a measure of "general intelligence," "mental ability," or any other of the loose, vague terms which are in current usage? How is the authority given, which enables us to take such a set of facts as a measure of specific ability? A moment's reflection will show that such authority is necessarily social. Now I do not think that a clear case can be made out for any of the criteria of intelligence which have been offered. The ordinary man is more likely, I think, to look upon excellence at school as the sign of special ability, rather than of general intelligence. And I question whether the *average* teacher working under *average* conditions is a competent judge of anything save average ability.

On the other hand, academic marks are surely a measure of ability of some kind, and it is equally certain that it would be better to say—abilities. In other words the obtaining of marks is a complex achievement. Unless we accept the opinion of the "man on the street" or of the *average* teacher, we must analyze this achievement into its causal factors, if we wish to gain a better understanding of what it does stand for. Now such an analysis can be made in terms of tests. That is, the functions which are active in obtaining marks can be expressed in terms of the functions which are active in making scores in tests. And *after* such an analysis has been made, we shall be able to approach the problem of interpreting both the test and the criterion to better advantage.^{2a} In a later portion of this paper the results of such

^{2a} The advantage gained by such an analysis may not be obvious to some readers. I trust it will become so as we proceed. At present I will say that

an analysis will be presented with the purpose of showing what it really is and how it enables us to approach nearer to the problem of determining the meaning of tests jointly with the analysis of the criterion. At present we must give our attention to the method of analysis. I may say at once that it is the method of partial correlation and that it is largely for the purpose of calling attention to its possibilities that the present paper is written. But as this method is relatively unfamiliar, and inasmuch as my chief purpose is the arousal of interest, I have thought it best to keep the methodological discussion informal and non-technical.

The position of the writer, then, up to this point, is that the end we should have in view in correlating tests with a criterion is the analysis both of the test and of the criterion. He protests against the naïf assumption of the mental tester that the test is, *ipso facto*, of a definite, more or less simple mental function. And he protests even more strongly against the more sophisticated, and hence more dangerous contention of the psychologist that the criterion, even though it is complex and vague in its significance, should be given artificial precision and simplicity *by definition*. The need is analysis. But it is a great deal easier to make this demand than to satisfy it. The difficulties with which the analyst has to deal lead us to considerations of another kind.

II. DISCUSSION OF THE METHOD

The first difficulty which we encounter is wholly artificial. The statement is often made by writers with a tender metaphysical conscience that a coefficient of correlation tells us nothing of true causal relations. Be that as it may, it does tell us quite as much about them as any other quantitative statement aiming at representing relations. For example, $v = gt$, such writers would say, tells us nothing of the nature of gravity. It informs us merely that, on this earth, the velocity of falling bodies varies

I conceive the process of acquiring knowledge of the symptomatic worth of tests as a *growth*. If we desire logical demonstration, we must put into our definitions what we mean to take out of them.

directly with the time, and that 'g' is the increment of velocity corresponding to an increment of time. Now, as we shall see, $v = gt$ is nothing but a regression equation. We can substitute for 'v' performance at the criterion, for 't' performance at the test, and for 'g' the coefficient of regression, and we are furnished with information which is analogous, though not perfectly so, to the case of falling bodies. One difference is that in the one case further interpretation of the "law" is centuries old and familiar, and it therefore seems obvious and simple. In the other case we do not, in most cases, have any very plausible interpretation which goes beyond the facts.

We encounter a more serious difficulty in the following. The statement is frequently made that the coefficient of correlation is meaningless in the case of non-linear regression.³ If this be true, the usefulness of correlation as a means of psychological analysis is seriously curtailed, for cases of *strictly* linear regression are rare. The "proof" of linearity simply means that non-linearity cannot be proven.⁴ It does not and cannot show that a straight line is the most probable regression, unless the line, (or lines), which passes through the means of the arrays is (are) actually a straight line. (The reader who is unfamiliar with this terminology is asked to reread this passage after he has read the next few pages.) Linearity is then assumed on account of its practical workability. But, fortunately, even when non-linearity *can* be proven, the statement is not true that the coefficient of correlation becomes meaningless. It merely loses some of its meaning. It is one of the great contributions of Yule to have shown the precise significance of the coefficient of correlation under any and all circumstances. This leads us to the consideration of points of a more technical kind.

In what follows it is assumed that the reader has a slight degree of familiarity with the terminology and the mathematical theory of correlation. Although the discussion is elementary, it does not aim to be an elementary exposition of the subject.

³ Brown. *Essentials of Mental Measurement*. Pp. 44-45.

⁴ The test for linearity is a farce with say one hundred observations, for the value of η can be made to vary within wide limits, for a single set of data, by varying the magnitude of the class-interval.

Still less does it pretend to give mathematical proofs. The reader who wishes to go into that phase of the subject is referred to the literature, and specific references will be made in their proper place. The aim is merely to present as simply as may be the points which the writer conceives to be essential for his purpose. Indeed if it should turn out that the reader who has had no acquaintance at all with the subject can follow the argument, the writer will be gratified. On the other hand I wish to guard very carefully against creating the impression that I am trying to pose as an expert mathematician. I conceive myself to have barely enough proficiency so that I *can* seize on some of the essentials and attempt some manipulation of a simpler sort. I am particularly anxious to make this acknowledgment in view of a number of instances where it should be made and is not.

Let us now suppose that a number of observed concomitant variations of two variables are expressed as deviations from their respective means and are then plotted, using rectangular coordinates, the means of the variables being at the origin of coordinates. Let us suppose, furthermore, that the points so plotted form a smooth continuous curve of some kind. This is a state of affairs approximated in the exact sciences. If now we are able to determine the equation of this curve, it is clear that we have in such an equation an expression which portrays accurately the amount of concomitant variation which actually occurs. For example, if we were to plot the results of a number of observations on the distance which a freely falling body covers during varying durations of time, we would find that the points would form a parabola, whose equation would enable us to estimate distance from time and vice versa. Suppose now that instead of a smooth curve our points form a jagged irregular line, and that it is required to find the straight line which most closely approximates the actual line. This may be done by the method of least squares, *i.e.*, upon the condition that the sum of the squares of the deviations from the straight line be a minimum. It may be shown that the line which satisfies this condition passes through the origin of coordinates. Its equation therefore will be $x = by$, where b is the tangent of the angle which the straight

line makes with the y axis. This equation may then serve *approximately* the same purpose as the equation for the parabola in the case of falling bodies. It represents, with *absolute* accuracy, the *average* amount of concomitant variation exhibited by our data. And if its use yields us a set of individual values which approximate the truth to the point of useful approximation, "regression" is said to be rectilinear. (Needless to say, this is not intended to be a *technical* definition of regression. I believe however that it gives a faithful account of its meaning).

Any curve or line which "fits" the data to the point of useful approximation is called a curve of regression. Such a line satisfies the condition of least squares for some given type (shape) of line. It is often said to be the line which passes through the means of the columns or rows respectively. Such a line obviously satisfies the condition of least squares. As has just been said, if it is straight, regression is said to be recti-linear or, briefly, linear. Its equation is the regression equation. Its tangent with the appropriate axis is the coefficient of regression. There are, of course, two such lines, one passing through the means of the x arrays, the other through the means of the y arrays.

We are not at all concerned, for our present purposes, with ways and means of evaluating the coefficient of regression. It is important to point out that the coefficient of correlation and the coefficient of regression are identical in value, when the deviation of the variables from their means are expressed in terms of their respective standard deviations as the unit of measurement. It follows that the *meaning* of the coefficient of correlation, in so far as it is a means of diagnosis and a measure of relation, is exhausted by the meaning of the coefficient of regression. Indeed it is possible to look on the coefficient of correlation as a convenient algebraic expression which enables us to find the value of the regression coefficient, to establish its validity or the degree of confidence which may be given to it, and to show the magnitude of error which may be looked for when it is used for prediction and diagnosis.

With this in mind let us now enter upon the closer examina-

tion of the claim that the coefficient of correlation is meaningless in cases of non-linear regression. In the first place we may admit without argument that the coefficient is meaningless with respect to the *form* of the true relation existing between the variables. If it is desired to describe the form, there can be no possible meaning in describing say a parabola by means of a straight line. Next we may note, without formal proof, that the coefficient will of necessity have a lower value than an expression which measures the amount of deviation from a curve of *closer* fit in analogous terms.⁵ It follows that, taken merely as an indication that an actual relation does exist between two variables, r , the coefficient of correlation, is actually entitled to increased confidence if non-linear regression is shown. Indeed the mere proof of non-linear regression is in and of itself proof of the existence of a true relation, and also of the fact that it is greater than indicated by r . It can hardly be claimed that a positive assertion which errs only on the conservative side is meaningless. As a special case we may note that $r = 0$ does not necessarily indicate the absence of relation.

Again, in the case of strictly linear regression, the errors of estimate will be equal for every part of the line and will be symmetrical as to sign.⁶ In the case of non-linear regression this will not be the case. For example, if the "true" regression is a sine-curve, the errors of estimate will be least at the points of intersection with the straight line of best fit, and will tend to be of opposite sign at different parts of the line. In short, the errors will tend to be systematic. But, when all is said and done, the straight line of best fit is what it claims to be, and, in the case of more than two variables, predictions and analyses made by its use will, on the average, be closer to the truth than any other conclusion based on the data and arrived at by any of the practically possible means. The subject of non-linear regression for the psychologist amounts simply to this. If he is investigating the relation of two variables to each other he can get nearest to

⁵ Such an expression is the correlation ratio, η . It does not, however, give us the equation of the curve of closer fit and hence is of no use for diagnosis.

⁶ The statement is true only for the types of regression usually found.

the truth by "fitting" a curve and determining its equation. Even in that case useful results are practically always obtainable by assuming linearity. But if one is dealing with a complex situation the only practical possibility with our present technique is to assume linearity. The results, when properly interpreted, will *not* be meaningless.

Suppose now, to take a hypothetical example, that we wish to ascertain the relation between crop yield and water supply in connection with a proposed irrigation scheme. We have available for the purpose data on the concomitant variation of rainfall and crop yield, and we find the coefficient of correlation of these two variables to be 0.40.⁷ We are then able to estimate the most probable crop yield we may expect from supplying a certain amount of water per acre. We will also be able to estimate, to any desired degree of probability, within what limits the actual yield will fall. In the case given, for any respectable degree of probability, the limits will be so wide as to render the information of little practical use. Now it may be argued that the relation between water supply and crop yield in a climate in which the sun is always shining is different from the relation which exists where water supply and sunshine very probably are inversely related. We wish to know the relation between yield and moisture supplied when there is a constant amount of sunshine, for this relation may be very much closer. If now in addition to our other data, we have data on the concomitant variability of "sunshine," we shall be able to supply this information. For suppose that we use the regression equation describing the relation between sunshine and yield to estimate the yield per acre. We estimate say 60 bushels of wheat. Actually it turns out to be 30 bushels. The difference then is an error of estimate. But, as we shall see at once, it is more than that. For consider that the correlation which we found, however low or high its probability, is in any case more probable than any other value known to us at present. Were it the "true" value, the error of estimate would be due exclusively to causes other than variation in the

⁷ The illustration, greatly modified and expanded, is taken from Yule's *Introduction to the Theory of Statistics*.

amount of sunshine. *It would, in fact, be an exact measure of the total effect of all operative causes, with the effect of sunshine eliminated.* As it is, it is just such a measure to the highest degree of probability possible to us with our present technique, and on the basis of the data at hand. We may therefore call an error of estimate a residual. Similarly, if we estimate the amount of rainfall during a given period from the amount of sunshine during that period, the residual will be a measure of the amount of rainfall associated during the period with all facts other than the somewhat obvious one that the sun is not shining when obstructed by clouds. (Obvious as it is, the fact is not irrelevant. At any rate, the reader is asked to fix his attention on the fact that a residual is a measure or representation of the association which exists between the fact we wish to estimate, and all other associated facts except one.) If now we actually compute all of the residuals which represent the relation of yield to all facts other than sunshine, and also all of the residuals representing the relation of rainfall to all facts other than sunshine, we shall have two sets of measures strictly analogous to our original data on the concomitant variation of yield and rainfall, except that the effect of sunshine has been eliminated from both measures. Now, if we compute the coefficient of correlation from these data, we shall have a measure of the relation which exists between yield and rainfall, the effect of sunshine being eliminated. In the notation of Yule,⁸ if crop yield is X_1 , rainfall X_2 , and sunshine X_3 , we shall have the value of $r_{12 \cdot 3}$.

⁸It may be desirable to give some account of this notation, sufficient to render its use in the text intelligible. All variables are denoted by subscripts of X , such as X_1 , X_2 , etc. The coefficient of correlation between any two is indicated by writing their subscripts beneath the symbol r , e.g. r_{12} . The coefficient of correlation between two variables, after the influence of other variables has been eliminated, is called a coefficient of partial correlation, and is written as follows. Let the variables whose relation is being expressed be X_1 and X_2 , and let the variables which have been eliminated be X_3 , X_4 , X_5 --- X_n . Then the coefficient of partial correlation is written $r_{12 \cdot 345 \dots n}$. The subscripts denoting the variables whose relation is being expressed are called "primary" subscripts and are written to the left of the point. Those denoting the eliminated variables are called "secondary" and are written to the right. r_{12} is called a coefficient of zero order, $r_{12 \cdot 3}$ is a coefficient of the first order, $r_{12 \cdot 34}$ of the second order, etc. In general, the

To be sure, this is not the method which is actually used. If it were, the amount of arithmetic would be almost infinite in complex cases. But Yule has shown that the method which is used is equivalent to such a method both in meaning and result.⁹ And, to my mind, this fact shows in the clearest fashion the *meaning* of partial correlation. Indeed it is largely for the sake of this single point which shows, I think, the simple causal reasoning, the simple logic, which underlies the complex mathematics which befuddles us, that the previous discussion has been given.

It follows that every claim we have made for the coefficient of correlation when only two variables are taken into account, is valid for the partial coefficient. That is, (1) Its meaning in cases of non-linear regression is clear and definite, (2) Its validity or "significance can be computed in the ordinary way, (3) The probable magnitude of average error incident to its use can be computed in the ordinary way. Indeed, as Brown says and says truly, "The full significance of correlation in psychology is to be found in the general theory of multiple correlation, of which the correlation of two variables is only a special case."¹⁰

We may express the relation of crop yield to both rainfall and sunshine in a single equation by simply adding the yield due to rain with sunshine constant to the yield due to sunshine with rain constant. Obviously such a sum is due to the combined effect of the two. That is, $x_1 = b_{12.3} x_2 + b_{13.2} x_3$. This is a partial regression equation. It would be represented graphically by a plane. Now the differences between the values of x_1 obtained

order of a coefficient denotes the number of secondary subscripts. The coefficient of multiple correlation, the meaning of which will be discussed below, denotes the relation which exists between a single variable, and the results which are obtained by estimating the values of that variable from a number of others by means of the regression equation. Its symbol is R , and is not to be confused with Spearman's R . The single variable is called "dependent," the others "independent." R is written $R_{1(234\dots n)}$ where 1 is the subscript of the dependent variable, 2, 3, 4, etc., the subscripts of the independent variables.

⁹ G. U. Yule. Proc. Roy. Soc. Series A, vol. 79, 1907. "On the theory of normal correlation for any number of variables treated by a new system of notation.

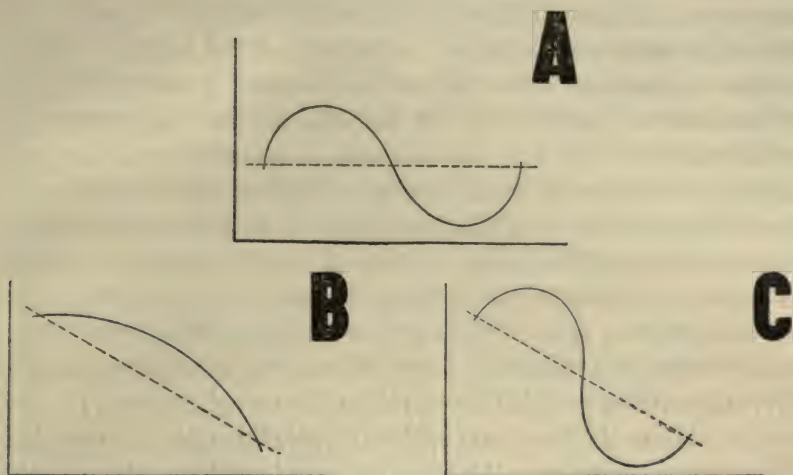
¹⁰ Wm. Brown. Essentials of Mental Measurement, p. 128.

by the use of this equation, and the actual values of x_1 , will be residuals containing the part of x_1 not associated with *either* x_2 or x_3 . These residuals might in turn be used to find the relation of a fourth variable to x_1 , with x_2 and x_3 eliminated, and so on indefinitely.

As has been said, the correlation which exists between the actual values of x_1 and the values estimated from an equation of partial regression is called multiple correlation, and the symbol of its coefficient is R . It is a measure of the closeness with which x_1 can be estimated from x_2 , x_3 , etc. It has some very useful properties which are important for the purposes of this paper and which we will discuss later.

Before leaving this part of the discussion, I feel that I am under moral obligation to call the attention of the reader to a statement by the highest authority on the theory of correlation, Prof. Karl Pearson. "The method (multiple correlation) . . . does assume that linearity applies within the degree of useful approximation. . . . The general linearity ought to be tested in all cases. Nothing can be learned of association by assuming linearity in a case with a regression line like A, much in a case like B."¹¹ (See diagram, page 15.) I wish to say that I realize fully my audacity—perhaps impertinence would be a more fitting phrase—in commenting on this statement. I am perfectly willing to follow blindly any course indicated by Karl Pearson. But, I am equally willing to do this when Mr. Yule leads the way. Now it does not seem to me that there is any *real* conflict of authority here and my interest in the subject compels me to point this out. It seems to me that the issue hinges on the meaning of the phrase "point of useful approximation." Mr. Yule has shown, and so far as I know his proof has not been challenged, that r retains an average significance under any and all conditions having to do with regression. If this be useful, the assumption of linearity is legitimate provided only an average significance is attached to the result. In a case like A there *isn't* any linear association. But if now the average slope

¹¹ K. Pearson. *Biom.* vol. 8, 1911-1912, p. 439. "On the general theory of the influence of selection on correlation and variation."



of the curve be changed (see diagram C) there will be such association and r will be its measure. The absence of association, on the average, was due to its real absence, *not* to the form of the regression. Whether or no such results are "useful" will be determined by the special conditions of the particular problem in hand. At any rate, they have a definite meaning.

Before leaving the topic, let me repeat what I said at its introduction. I have tried throughout to make clear the meaning of certain phases of the topic which I deem essential for my purposes. I have not tried to prove anything. The reader has been asked to accept the statements made on the authority of Yule.¹²

What now is the significance of partial correlation for the Mental Test situation? In the first place it furnishes us with the only means at present available for the analysis of the test and the criterion. For suppose that we find a correlation of $+0.28$ for the Logical Memory test with academic marks. The previous discussion should have made it clear that this is *not* a measure of the extent to which "logical memory" is a factor in "general intelligence." As any instructor will testify, marks are not even a very good measure of academic intelligence, and a

¹² In addition to references given above, See G. U. Yule. Proc. Roy. Soc. Vol. 60, 1897. "On the significance of Bravais' formulae for regression, etc. --."

cursory examination of the test in say Whipple's Manual will show that most of the logic is contained in the name. This is further emphasized when the test is evaluated quantitatively, as it must be for purposes of correlation. Mechanical memory is by no means confined to the learning of nonsense syllables and digits. If it were, it would not be a problem for education. I can conceive of at least the possibility of so organizing a nonsense syllable test that, as a criterion of *logical* memory, it would be superior to the test which bears that name. At any rate, it is not all certain, *a priori*, that a student who depends largely on a rather verbal and mechanical type of memory is at a serious disadvantage either in this test or in the matter of marks. If now we are able to devise a test which, *a priori*, carries a somewhat stronger presumption of *logical* memory, we have at any rate *some* material for analysis. For let X_1 stand for Marks, X_2 for the original logical memory test, and X_3 for the supposedly more logical test, then $r_{12 \cdot 3}$ carries a stronger imputation of being rather verbal than r_{12} , and there is a greater probability that $r_{13 \cdot 2}$ stands for a more logical type. For that which the two have in common has been eliminated in each case in so far as it is associated with marks, and therefore the relation to marks of that which is peculiar to it stands out more clearly. We see that we have here the promise of a fruitful way of combining experimental and statistical research. For a slight modification of the test, or even a change in the method of scoring, may lead to results of significance with regard to the nature of the test. This will be illustrated later. Speaking generally, a partial coefficient ($r_{12 \cdot 34 \dots n}$) is more easily interpreted than a coefficient of zero order (r_{12}), for in the case of r_{12} we have to face the vague question why X_1 is related to X_2 , while in the other case we may ask what in X_2 is related to X_1 that is peculiar to it, and has nothing to do with X_3 , X_4 , etc. In other words, we have more data on which to base analysis.

Another use to which partial correlation might be put is in connection with so-called "practical" diagnostic work. If a number of tests have been given, it is usually desired to combine them into a single measure of their diagnostic value, with reference

to a single criterion such as marks. The method sometimes used is that of expressing all the scores as deviations from their mean with their respective standard deviation as the unit of measurement. The scores for each test for each individual are then added, and these combination scores are correlated with the criterion.¹³ The method is somewhat faulty. It attaches equal importance to all tests regardless of their correlation with say marks, and it is perfectly obvious that this is false. Worse, when handled wrongly, it may even serve to conceal linear relations easily discernible in the data. For, to take an extreme case, suppose we combine $r_{12} = + 1.00$ and $r_{13} = - 1.00$ according to this method. The result, if the two distributions happen to be nearly parallel, will approximate zero, and we shall have succeeded in converting two perfect diagnostic tools into an absolutely useless one. This source of error can be obviated by computing all correlations of type r and reversing the signs of the scores of all tests showing a negative correlation, reversing the meaning to correspond. At its best, the method is purely empirical and the meaning of the coefficient obtained by its use is neither clear nor definite (except mathematically). We cannot reason from it, we cannot use it as an analytic tool. If it does not possess diagnostic value, it possesses no value whatever.

The best method of combining a number of tests is to find their regression equation. R , the coefficient of multiple correlation, will be the indirect measure of the diagnostic value of this equation. As we shall see, R is exceedingly valuable for analysis, is more easily calculated than the regression equation, and, unless we wish to apply diagnosis to the case of single individuals, it renders the finding of the regression equation unnecessary. In so far as he knows, the use of R for such purposes is original with the writer. When regression is truly linear, R gives us, within the limits of accuracy of sampling, a measure of the actual relation which exists between two complex set of facts, the criterion and the tests. Besides it will enable us to make an accurate analysis of the criterion in terms of the tests, in so far as it is associated with the tests. When regression is non-linear,

¹³ R. S. Woodworth. *Psy. Rev.*, 1912, p. 97.

the results have the same significance to a lesser degree of approximation.

In view of all this, it may seem somewhat surprising that the method has not found more frequent application. One reason for this is perhaps that the subject is somewhat difficult and not well understood generally. Another reason is probably the large amount of arithmetical labor called for in complex cases. Yule states that the working out of a case involving eight variables is practically beyond the powers of a single individual.¹⁴ Kelley¹⁵ states that in the case of eight variables it is practically necessary to resort to an approximation.¹⁶ But Kelley himself has reduced the amount of mechanical labor materially through the publication of a very useful set of tables, contained in the bulletin referred to. I myself have devised a scheme of procedure, involving the use of R, which makes it possible to reach all of the results which Kelley reaches, and perhaps a little more, with approximately half the work indicated by him. A full exposition of this schema will be found in the appendix. These mechanical improvements, as well as the fact that a complete working out of all possible relations is not necessary, bring the method within the reach of a single individual. For the purposes of the present paper I worked out a case of sixteen variables without resorting to approximation. It took me a little over two months, but I did a lot of useless work. I could do this work now in five or six weeks at the most. Of course I was working with a comparatively small number of observations, but, after the coefficients of zero order have been found, the amount of arithmetic does not depend on the number of observations. But be that as it may, the method *must* come into use if scientific analysis is ever to take the place of blind fumbling about. For it is the only method

¹⁴ G. U. Yule. Roy. Stat. Soc. Journ., vol. 60, p. 182.

¹⁵ Kelley is the only American psychologist who has exploited the method of partial correlation. See his "Educational Guidance," Teachers College, Columbia University, Contributions to Education, No. 71. In the opinion of the writer, Kelley's work loses some of its value through his failure to call attention to some of the difficulties and sources of misunderstanding with which the subject is hedged about. These difficulties are still in front of us.

¹⁶ Bull. No. 27, U. of Tex., May 1916, p. 18.

available at present that holds out even a hope of making systematic progress in attacking situations as complex as those with which we have to deal.

For the present however let us dwell rather on the limitations of the method. In the first place it would be quite erroneous to suppose that the magnitude of R , and consequently its value, can be indefinitely increased by simply increasing the number of tests. To take another illustration from Yule,¹⁷ if $r_{12} = 0.8$, $r_{13} = 0.4$, and $r_{23} = 0.5$, it would be quite natural to suppose that X_1 could be estimated with greater accuracy from both X_2 and X_3 , than from X_2 alone. But this would be quite wrong, because $r_{13 \cdot 2} = 0$. In other words, everything in X_3 that has diagnostic value, or that is associated with X_1 , is contained in X_2 . Pearson dwells at length on this point.¹⁸ For example, if an infinite number of variables are correlated equally with each other, the value of r being 0.5, he shows that R with reference to any one of them is 0.71. In the case of ten such variables, $R = 0.67$, for five variables, $R = 0.65$. The difference in diagnostic value between 0.65 and 0.71 is negligible. How serious a difficulty this is will appear later. At present let us note that the trouble is not with the method, but with the material with which it deals.

The other difficulty to which I wish to call attention is one of interpretation. R is essentially positive regardless of the signs of the coefficients of the regression equation. It is therefore subject to biased error, that is, errors due to fluctuations of sampling will not tend to neutralize each other, but will be cumulative. Consequently the "probable error" of R is not a true measure of its validity. It should be compared to the value of R in the case of a number of really uncorrelated variables owing to fluctuations of sampling alone. Pearson, in the article referred to below, promises us a formula for finding such a value, but, to date, I have not found it in the literature accessible to

¹⁷ Intro. to the Theory of Statistics, p. 237.

¹⁸ K. Pearson. *Biom.* vol. 10, 1914-15, p. 181. "On certain errors with regard to Multiple Correlation occasionally made by those who have not adequately studied the subject."

me.¹⁹ Yule²⁰ publishes an approximation formula which gives the value. It is $(n - 1)^{1/2} / N^{1/2}$, where n is the number of variables, and N the number of observations. Inspection of this formula will show that its value may easily become unpleasantly large. For example, if $n = 16$, and $N = 92$, as is the case in the problem I worked out in the present paper, then $R = 0.40 \pm 0.06$. That is, although R is seven times as large as its probable error, it has no validity whatsoever. It is the failure to call attention to this fact which I alluded to in discussing the work of Kelley.

III. APPLICATION OF THE METHOD TO CORRECT DATA

In view of these two serious limitations I might perhaps be asked why I undertook so laborious a task as the computation which forms a part of this paper. The answer is simply that it is no part of my purpose to advertise a method for making a silk purse out of unsuitable material. There is in this paper no attempt to break the record for altitude. The value of R does not interest me except in so far as it is an instrument of analysis. My object was to sift a mass of typical material down to its significant constituents. The general character of the results I was fairly sure of before I undertook the work, but I could not know precisely what material would be retained by the meshes of the sieve. I could have reduced the number of variables to eight or nine and have been morally certain that I was not discarding anything of value, but it is questionable what weight my moral certainty would have carried with others. Besides I fancied that a drastic concrete illustration of the difficulties which I have just called attention to would do no harm. Furthermore I am very much interested in the subject of partial correlation and hope that the present paper, in conjunction with the appendix, will have a methodological value. The particular material was used simply because, owing to the kindness and scientific attitude of Dr. Kitson, it was available. But tests are not the only things capable of analysis. Nor are they the only source of information on which diagnosis can be based. For example, if Kelley's work

¹⁹ Pearson, *Biom.* vol. 8, p. 437, *op. cit.*, p. 18.

²⁰ *Proc. Roy. Soc.* 1907. *Op. cit.*

were based on more than thirty-three observations, it would show that the six tests which he uses and combines with academic marks in grammar school and with judgments of ability by the teachers—are negligible for the purpose of predicting performance in High School in comparison with these other means of diagnosis.²¹ However I ought to say that up to a time when the work was more than half finished I thought I would have a considerably greater number of observations at my disposal than proved to be the case. This brings us to the consideration of the material with which the present investigation deals.

This material was put at my disposition by Dr. Kitson. Neither he nor I dared to hope that the investigation would lead to results of final validity. In the first place the number of observations were not adequate. Besides Dr. Kitson himself looks upon the stage of his work from which the data were taken as its pioneer stage. Since that time he has added new tests and has improved the old tests. If it had been possible to include this later material, there is every reason to believe that more significant results would have been obtained. I do believe that needless duplication of identical functions is a feature of *all* lists of tests in actual use if they are at all extensive. In order to enable me to test this out, Dr. Kitson furnished me with such material as he had available. Even though the results do not have a direct bearing on his later work, or on similar work done by others, it was thought that the indirect light they would cast would have value.

The material then is the same as that obtained by Dr. Kitson for his "Scientific Study of the College Student."²² As may be seen by referring to this monograph, Dr. Kitson gave a large number of tests to the students of the college of Commerce and Administration at the University of Chicago. The work there described covers a period of two years. It has been continued and the results of two more years are now available. At the time my own work was done, the academic marks for the fourth year were not at hand. Besides so many changes had been made

²¹ T. L. Kelley. Ed. Guid., p. 71 ff.

²² Psy. Rev. Mono., 1917, vol. 23, No. 1.

in the tests themselves that their combination with the other three years did not seem feasible. There did not, however, appear to be any *a priori* reason why the first three years could not be combined, so that I expected to have 150 sets of observations on which to base any conclusions I might reach. I did not investigate the three years separately until, when the work was over half finished, certain differences forced themselves on my attention. Then I did investigate and found, amongst other differences, that the academic marks for the third year differed significantly from those of the other two years. Inquiry amongst the members of the faculty resulted in conflicting evidence as to the reasons for this, so that I was compelled, regretfully, to discard these data. This reduced the number of observations, or subjects, to 92, and the period covered became precisely the period described in Dr. Kitson's Monograph, *i.e.*, the academic years 1913-14, and 1914-15.

In this group there are included 80 freshmen and 12 sophomores; 39 freshmen and 6 sophomores in 1913, and 41 freshmen and 6 sophomores in 1914. The desirability of the inclusion of the sophomores may be questioned. However it is the practice at the School of Commerce and Administration to test all students who have not previously had the tests, and in such a group there always are a certain number of individuals who come from other departments or institutions with advanced standing. Consequently such a group is representative in a definite sense.

As the Psychological Review Monographs are quite accessible, I deem myself absolved from the uninteresting task of copying the full description of these tests. Most of them are standard and well known. Also their names give a good indication of their general nature. As much additional description as seems essential will be given informally with the discussion. At present I give only the names in conjunction with the numbers which they were given in the present study, and a brief description of the more important tests. The list follows. As will be seen, the criterion, academic marks, is given No. 1.

(1) Academic Marks.

(2) Immediate memory for logical material, heard.

- (3) Immediate memory for logical material, seen.
- (4) Loss or gain of log. mat., heard, after two weeks.
- (5) Loss or gain of log. mat., seen.
- (6) Sentences Built.
- (7) Hard Directions, printed, speed.
- (8) Constant Increment, speed.
- (9) Memory for objects seen.
- (10) Number-checking (cancellation).
- (11) Opposites, speed.
- (12) Memory for numbers heard, (span).
- (13) Word building.
- (14) Opposites, accuracy.
- (15) Constant Increment, accuracy.
- (16) Hard Directions, accuracy.

The description of the seven tests which will be of most interest to us follows:

No. 2. Logical Memory, immediate, auditory.

Materials: Blank sheet of paper and pencil.

Directions: "I am going to read you a rather long passage and shall ask you to listen very carefully, for when I have finished I wish you to reproduce the meaning of the passage. The passage is too long for you to remember word for word, but try to get the entire meaning, then in reproducing, use the same words as appear in the text whenever you can.

The Passage: The passage may be characterized as popular science.

Method of Scoring: It will be noted that this passage contains a main proposition and three illustrations, the last one of which is amplified. For reproduction of the main proposition, two units were given; for mention of the first, second, and third illustration there were given 14, 13, and 14 units respectively. Thus by merely stating the main proposition and the illustrations the individual could score 43. In addition to these gross divisions, the passage was further divided into 81 ideas. Counting each one of these as two thirds of a unit, their united value is 54, which added to the 43 unit mentioned, permits scoring on a basis of 97 points for correct reproduction of the passage."

No. 4. Logical Memory, immediate, visual.

Materials: See directions.

Directions: "On the reverse side of the paper before you will be found a long passage which I wish you to read carefully when I give the signal. Read it but once, then turn it over, and on the back of it write all you can recall of the passage. Be careful to read each sentence but once, then turn over the paper and reproduce the meaning as accurately as possible."

The Passage: May be characterized as popular psychology.

Scoring: Same as in No. 2.

No. 3. Loss or gain, Logical Memory, auditory.

Direction: Write all you can recall of the passage I read to you at the last psychological examination, beginning "More than once it has happened in the history of science."

Scoring: The papers were first scored as in two. Then the difference between No. 2 and No. 4 was taken as the score of No. 4.

No. 5. Loss or gain, Logical Memory, visual.

Analogous to No. 3 in every way.

No. 9. Memory for Objects, visual.

Materials: Covered box twelve by twenty-three inches, containing the following objects fastened to the bottom: fountain-pen, pencil, twenty-five cent piece, envelope, inkwell, maroon ribbon, ruler, pen-filler, two-cent stamp, and key.

Directions: I am going to show you a group of objects for six seconds, then will ask you to name them aloud from memory.

Scoring: The score represents the number of objects correctly reproduced.

No. 6. Sentences Built.

Directions: I will give you five minutes in which to make as many sentences as possible containing three words which I will give you presently. For example, if I gave you the words money, river, Chicago, you might make a sentence like this: "Chicago spends much money improving its river." You may use either singular or plural forms of the words, nominative, objective, or possessive case. Simply use all three of the words in a sensible sentence and make as many different sentences as possible. The three words are,—citizen, horse, decree."

Scoring: The score represents the number of sentences formed.

Extract from Dr. Kitson's comment: "... the papers which contained a relatively large number of sentences necessarily showed much sameness in subject matter and structure."

No. 8. Constant Increment.

Material: Card containing one hundred two-place numbers.

Directions: I am going to give you a list of 100 numbers and shall ask you to add four to each number as quickly as possible, giving the sum aloud. You may practice on this list: 22, 34, 92. Begin at the top of each of the four columns and add four to each number. You need not be afraid to go fast for the test is easy and you are not likely to make mistakes. You should be accurate, however, because every error will take off one point from your score. The main thing is to add as rapidly as possible.

Scoring: The number of errors was the accuracy score. The number of seconds, the time score.

The materials accessible to me were the gross scores in all of the fifteen tests. In no case were the original records at hand, as they had been destroyed some time before. Had they been avail-

able, it would have been possible to push analysis further than was actually the case, and many of the suggestions made below could have been investigated. I was not however obliged to accept any empirical indices, as all indices which Dr. Kitson used were based on the gross scores.

This, practically, is all the information we need for the present. Now our problem in evaluating these tests is not simply that of determining their total relation to marks. If it were, we could solve it easily and directly by simply computing the fifteen correlations which indicate this relation. We should find, for example, that the correlation of the Auditory Logical Memory test with marks is $+0.28$, and that of Hard Directions with marks is $+0.25$, or using the numbers assigned to these tests, $r_{12} = +0.28$, $r_{17} = +0.25$. But such numbers would tell us nothing of the nature of the relations in each case. The functions measured by r_{12} and r_{17} might in reality be identical, independent, or even mutually exclusive. There would be little sense in debating such an issue on *a priori* grounds when we can find directly that $r_{27} = +0.30$. We now have the beginnings of analysis, for we know, to the degree of probability of which our data permit, that the two functions are different in some respects and identical in others. A similar line of reasoning might be applied to every one of these fifteen tests, or variables, paired successively with each one of the others. In table I there will be found a complete list of all possible correlations of our fifteen tests amongst themselves and with marks, 120 all told. They are given in full because, aside from the raw data which are too bulky to print, they represent the complete data for this study. But, aside from this, the reader will be able to convince himself by a study of this table that the data in their present shape are far too complex to permit of conclusions more definite than the very vague one we have just stated with reference to No. 2 and No. 7, *i.e.*, that they are alike in some respects and different in others. (Of course, if we have had practice with the method of partial correlation, we may be able to go a little further, for we would be able to guess with some accuracy the results of computations based on the data.) But some such problem as whether No. 2

has some characteristic peculiar to it, or whether it is exhaustively represented by the other fourteen, would be quite insoluble.

To answer such a question we must resort to partial correlation. Table 2 shows the effect of successively eliminating the effect of each variable, as we may conveniently call our tests, on the relation between No. 2 and No. 1.

TABLE 1

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1															
2	28														
3	17	44													
4	17	-19	-01												
5	10	06	-20	55											
6	26	05	14	-10	-07										
7	25	30	04	08	02	13									
8	21	03	08	00	-14	31	17								
9	-12	10	00	02	02	08	-03	25							
10	11	10	05	04	-10	-06	25	36	17						
11	10	41	04	06	09	10	17	23	14	01					
12	07	13	22	00	-01	09	22	01	-17	16	-09				
13	09	17	18	09	06	20	11	25	08	23	20	18			
14	-03	02	-08	-07	-10	11	-04	17	15	-18	18	-01	02		
15	08	03	-04	-06	-10	18	-04	40	15	04	07	01	03	12	
16	05	20	14	10	03	03	25	42	09	-04	21	13	10	16	-01

If the reader will now recall what was said about the association between "Residuals," the meaning of this table should be clear. For example $r_{12 \cdot 16} = + 0.28$ is a measure of the association of No. 1 and No. 2 in so far as neither No. 1 nor No. 2

TABLE 2

r_{12}	$= + 0.28$
$r_{12 \cdot 16}$	$= + 0.28$
$r_{12 \cdot 15 \cdot 16}$	$= + 0.27$
$r_{12 \cdot 14 \cdot 15 \cdot 16}$	$= + 0.27$
$r_{12 \cdot 13 \cdot 14 \cdot 15 \cdot 16}$	$= + 0.26$
$r_{12 \cdot 12 \cdot 13 \cdot 14 \cdot 15 \cdot 16}$	$= + 0.26$
$r_{12 \cdot 11 \cdot 12 \cdot 13 \cdot 14 \cdot 15 \cdot 16}$	$= + 0.25$
$r_{12 \cdot 10 \cdot 11 \cdot 12 \cdot 13 \cdot 14 \cdot 15 \cdot 16}$	$= + 0.24$
$r_{12 \cdot 9 \cdot 10 \cdot 11 \cdot 12 \cdot 13 \cdot 14 \cdot 15 \cdot 16}$	$= + 0.26$
$r_{12 \cdot 8 \cdot 9 \cdot 10 \cdot 11 \cdot 12 \cdot 13 \cdot 14 \cdot 15 \cdot 16}$	$= + 0.31$
$r_{12 \cdot 7 \cdot 8 \cdot 9 \cdot 10 \cdot 11 \cdot 12 \cdot 13 \cdot 14 \cdot 15 \cdot 16}$	$= + 0.28$
$r_{12 \cdot 6 \cdot 7 \cdot 8 \cdot 9 \cdot 10 \cdot 11 \cdot 12 \cdot 13 \cdot 14 \cdot 15 \cdot 16}$	$= + 0.28$
$r_{12 \cdot 5 \cdot 6 \cdot 7 \cdot 8 \cdot 9 \cdot 10 \cdot 11 \cdot 12 \cdot 13 \cdot 14 \cdot 15 \cdot 16}$	$= + 0.28$
$r_{12 \cdot 4 \cdot 5 \cdot 6 \cdot 7 \cdot 8 \cdot 9 \cdot 10 \cdot 11 \cdot 12 \cdot 13 \cdot 14 \cdot 15 \cdot 16}$	$= + 0.36$
$r_{12 \cdot 3 \cdot 4 \cdot 5 \cdot 6 \cdot 7 \cdot 8 \cdot 9 \cdot 10 \cdot 11 \cdot 12 \cdot 13 \cdot 14 \cdot 15 \cdot 16}$	$= + 0.32$

is associated with No. 16, (accuracy, 'Hard Directions'). But this value is identical with the one obtained before X_{16} was eliminated. It follows, even though $r_{2\ 16}$ is $+0.20$, that the relation of No. 2 to No. 16 is of no significance with reference to the relation which exists between No. 1 and No. 2. The converse, however, is not true. For $r_{1\ 16} = 0.05$ and $r_{1\ 16} \cdot 2 = 0$. Therefore the relation of No. 1 to No. 16 is entirely accounted for by that which No. 2 and No. 16 have in common. These conclusions are subject to the limitations imposed on us by the small number of our observations and by the assumption of linear regression. We have discussed the second, and we will presently discuss the quantitative expression of the first of these limitations.

Similarly $r_{12 \cdot 3\ 4 \dots 16} = +0.32$ is a measure of the relation of No. 1 and No. 2 in so far as both No. 1 and No. 2 are not associated with any of the other fourteen variables. This relation is peculiar to academic marks and to Immediate Logical Memory (Auditory), alone. This answers the question we propounded, and is analysis in a very real sense. But before we endeavor to push the analysis to its logical conclusion and try to ascertain the character of this elementary function which we have isolated, we must turn our attention to the disillusionizing subject of validity. We have deliberately avoided this topic so far, because its discussion would have added nothing to our understanding of the causal reasoning which underlies the theory of correlation. Now that we have reached concrete results it can no longer be postponed.

If a number of samples be taken from a universe of discourse (*viz.*, the universe of college freshmen of America), and constants such as the Mean, the Standard Deviation, and the Coefficient of Correlation be ascertained, the results will differ from what would have been obtained if the entire "universe" had served as a basis. Such deviations from the theoretical true value are called errors of sampling. Theoretically it is possible to ascertain the probability that any given error of sampling falls within given limits. Practically, in the case of r , this is done on the basis of the assumption that distribution is normal. The conventional expression given to facilitate the computation

of this probability is the so-called probable error. In and of itself, however, it merely means that the deviation of r from its "true" value is as likely to be greater than the true value, as it is to be smaller. If a given r is equal to its probable error, the chances are 1:1 that it would arise by chance even though the "true" value be zero. If $r = 2$ P.E. (probable error), these chances are 1:5, roughly. For $r = 3$ P.E., they are 1:23, for $r = 4$ P.E., 1:143, etc. —. When $r = 3$ P.E., r is said to be "significant." Of course, such a standard is conventional and arbitrary, and not all authorities recommend the same ratio. In any case it is well to bear in mind the meaning of this "significance." With these considerations in mind let us now return to the consideration of our data and results.

In table 3 column 1 (see below) are given the r 's of marks with each one of our tests, followed by the probable error of r . They are the coefficients of zero order. In column 2 the same values and their probable errors are given after the influence of the other fourteen has been eliminated by partial correlation. These are the coefficients of the 14th order. They represent the correlation with marks of what is unique to each test. Let us return to test 2 and 3 the logical memory test. $r_{12} = +0.28$, $r_{13} = +0.17$. Does this show that "Auditory Presentation" is more highly correlated with marks than Visual Presentation? Not at all, for the difference is 0.11 and the probable error of this difference is 0.094. So the chances are about even

TABLE No. 3

2. Log. Mem. Aud.	+ 0.28 $\pm$ 0.065	+ 0.32 $\pm$ 0.063
3. Log. Mem. Vis.	+ 0.17 $\pm$ 0.068	+ 0.04 $\pm$ 0.070
4. Loss or Gain in No. 2	+ 0.17 $\pm$ 0.068	+ 0.26 $\pm$ 0.066
5. Loss or Gain in No. 3	+ 0.10 $\pm$ 0.070	+ 0.03 $\pm$ 0.070
6. Sentences Built	+ 0.26 $\pm$ 0.066	+ 0.23 $\pm$ 0.067
7. H. Directions, Speed	+ 0.25 $\pm$ 0.066	+ 0.09 $\pm$ 0.070
8. Con. Increment, Speed	+ 0.21 $\pm$ 0.067	+ 0.21 $\pm$ 0.067
9. Objects Seen	— 0.12 $\pm$ 0.069	— 0.23 $\pm$ 0.067
10. Number-Checking	+ 0.11 $\pm$ 0.069	$\pm$ 0.01 $\pm$ 0.070
11. Opposites, Speed	+ 0.10 $\pm$ 0.070	+ 0.12 $\pm$ 0.069
12. Numbers Hard	+ 0.07 $\pm$ 0.070	$\pm$ 0.04 $\pm$ 0.070
13. Words Built	+ 0.09 $\pm$ 0.070	$\pm$ 0.07 $\pm$ 0.070
14. Opposites, Accuracy	— 0.03 $\pm$ 0.070	$\pm$ 0.01 $\pm$ 0.070
15. Con. Increment, Accuracy	+ 0.08 $\pm$ 0.070	$\pm$ 0.03 $\pm$ 0.070
16. H. Directions Accuracy	+ 0.05 $\pm$ 0.070	— 0.14 $\pm$ 0.069

that the difference is due to chance. But if we turn to the corresponding r 's of the 14th order, we find that $r_{12 \cdot 8 \cdot 4 \dots 16} = +0.32$ and $r_{13 \cdot 2 \cdot 4 \dots 16} = +0.04$. The difference is 0.28, the probable error of the difference 0.094. Hence the chances that this difference is due to fluctuation of sampling alone are 1:23. By conventional standards, there is a valid difference between Auditory and Visual presentation. *Now* we can interpret. The difference is due to something that has not been eliminated. I can think of but three possibilities. They are, (1) a difference in subject matter, (2) speed of presentation. In the visual presentation the subject reads at his own rate and may violate the instructions against re-reading. In auditory presentation he must accept the rate of speed of the experimenter. (3) A difference specific to the sense avenue, possibly in conjunction with the previous experience of the individual. Our data do not permit of a choice between these three possibilities, but it would be a simple matter so to control these conditions in another series of tests that interpretation would be narrowed down practically to a single possibility. Of course someone else might think of other possibilities. But he should remember that the cause of the difference cannot be anything involved in any of the other tests, unless it be a very marked difference of degree, and also that elimination has been from the criterion as well as from the test. (I have not thought it necessary to mention possibilities which would come under the head of obvious control of conditions common to all careful experimentation.)

Again, $r_{12} = +0.28$, $r_{14} = +0.17$. Recalling that 4 is the difference between the score in test No. 2 and the score made after two weeks, we may note that the factor of "immediate memory" has been eliminated not by partial correlation, but by the method of scoring. The change is so radical that it does not seem profitable to compare 2 and 4 as modifications of the same test. We may note, however, that r is not significant by conventional standards, but becomes so as a partial of the 14th order, for $r_{14 \cdot 23 \dots 16} = +0.26$. Although it is not apparent from table 3 alone, the same thing is true of tests No. 3 and No. 5. No. 3 and No. 5 are identical with No. 2 and No. 4 respectively ex-

cept as to subject matter and mode of presentation. Their corresponding values, of zero order, are $r_{13} = +0.17$ and $r_{15} = +0.10$. Their coefficients of the 14th order are 0.04 and 0.03. But we have already seen that Auditory presentation contains everything of significance in Visual presentation, so that there is nothing left to compare after No. 2 and No. 4 have been eliminated, as they have been in the 14th order. To make the comparison we must contrast the coefficients of the 12th order. We find $r_{13 \cdot 56 \dots 16} = +0.19$ and $r_{15 \cdot 56 \dots 16} = +0.20$, results which are quite similar in character to what we found in the case of No. 2 and No. 4.

Now the significance of No. 4 (and No. 5) has been covered up in some way. Consulting a table of some 1300 partial coefficients which were computed in order to get our results, but which are not published, we find that $r_{14 \cdot 2} = +0.24$. The same fact might have been guessed from $r_{24} = -0.19$ (see table No. 1). The significance of No. 4 then, is obscured by the fact that subjects who make a high score "immediately" tend to forget more than those who do not. No doubt this is at least in part due to the fact that they have more to forget. Of course, different results as to r_{24} might have been obtained if the loss or gain had been expressed in percentage terms. (E.g., if $X = 80$, X after two weeks $= 60$, then $X = 20/80 = 25\%$.) But this would have been as arbitrary as the method chosen. Besides, and the fact is interesting, using partial correlation made us relatively independent in the matter of selecting a unit of measurement, for if we had adopted "percents" and had obtained say $r_{24} = 0$ (instead of -0.19), we would also have had a different value for r_{14} and would have had the value of $r_{14 \cdot 2} = +0.24$, as before, unless other factors entered.

Returning now to the interpretation of $r_{14 \cdot 23 \dots 16}$, we have seen that it is significant and that it is a factor in r_{12} which not only was itself hidden, but also operated to lower the value of r_{12} . This factor cannot be subject matter, for that is the same in both cases. It cannot be "mode of presentation," for X_4 is not stated with reference to the material as presented, but with respect to the material as "immediately" retained. Besides the

original mode of presentation is the same in both cases. The factor may be that of interest in the tests themselves. Interested subjects will tend to rehearse their performances and will be likely to discuss the tests with others and to compare notes. Or else it is reasonable to argue that the subjects who have seized the essential significance of the passage will have an advantage in the matter of permanent retention over those who depend on a more verbal type of memory, whereas in test No. 2 as scored, only two units out of a possible 97 are allowed for the reproduction of the gist of the passage, and 54 units are allowed for the reproduction of "ideas" which may be nothing but words.²⁸ Again I am obliged to say that our data do not justify me in saying more. But it is obvious that a comparison of different methods of scoring would be likely to give a definite conclusion to our problem.

Let us now consider No. 8, the Constant Increment test. We note that $r_{18} = +0.21$ and $r_{18 \cdot 23 \dots 16} = +0.21$. Both r 's are significant, but it is difficult to draw a conclusion beyond the restatement of the fact that No. 8 contains a significant factor not contained by the other fourteen. As we shall see, even the 1300 coefficients referred to above do not contain the necessary information which would enable us to push analysis much further. But it is quite unthinkable that such a complex function as adding a constant increment under test conditions should resist analysis. Indeed analysis is possible on the same lines we have pursued so far. But now the amount of arithmetic necessary for analysis might easily become unthinkable. It is at this point that the method devised by the writer will enable us to proceed.

Table No. 4, given below, is somewhat analogous to table No. 2. It shows the effect on the relation of marks to No. 8 of the elimination of each of the other fourteen tests for some one order. It is unlike table No. 2 in that elimination is not consistently successive or continuous. The reason for this is practical. To have the order of elimination successive and continuous for

²⁸ Kitson, *op. cit.*, p. 25.

TABLE No. 4

r_{18}	$= + 0.21$
$r_{18} \cdot 16$	$= + 0.21$
$r_{18} \cdot 15 \ 16$	$= + 0.19$
$r_{18} \cdot 14 \ 15 \ 16$	$= + 0.20$
$r_{18} \cdot 13 \dots\dots\dots 16$	$= + 0.18$
$r_{18} \cdot 12 \dots\dots\dots 16$	$= + 0.19$
$r_{18} \cdot 11 \dots\dots\dots 16$	$= + 0.18$
$r_{18} \cdot 10 \dots\dots\dots 16$	$= + 0.17$
$r_{18} \cdot 9 \dots\dots\dots 16$	$= + 0.18$
$r_{18} \cdot 2$	$= + 0.21$
$r_{18} \cdot 2 \ 3$	$= + 0.21$
$r_{18} \cdot 2 \ 3 \ 4$	$= + 0.21$
$r_{18} \cdot 2 \ 3 \ 4 \ 5$	$= + 0.21$
$r_{18} \cdot 2 \dots\dots\dots 6$	$= + 0.13$
$r_{18} \cdot 2 \dots\dots\dots 7$	$= + 0.12$

each of the fifteen correlations of type r , would involve the computation of 8400 coefficients, instead of the 1300 which we have. (See appendix.) We note that $r_{18} \dots\dots$ fluctuates about its original value except at the point where No. 6 (Sentences Built) is eliminated, and there it drops to $+ 0.13$. I am unable to offer any very convincing interpretation as to the factor which is alike in these two tests. Instead we will face the question how $r_{18} \cdot 2 \dots\dots\dots 16$ regains, so to speak, its original value. Clearly table No. 4 does not furnish an answer. Neither do any of the 1300 coefficients from which it is an excerpt. The reason is as yet hidden somewhere in the infinite complexities of a situation which involves sixteen variables. Our only hope of finding this factor, without an inordinate amount of arithmetic, lies in reducing the number of variables. Now the factor, be it what it may, *must* be significantly related to marks. Therefore, if we can find the variables which, *when combined*, include *everything* which is so related, and if we can exclude those which merely duplicate some factor or set of factors, we will be that much nearer to the solution of the problem.

Let us recall that R , the coefficient of Multiple Correlation, is a measure of the relation of a number of combined variables to another variable. $R_{1(2 \ 3 \ 4 \dots\dots\dots 16)}$, the relation of our fifteen tests to marks, is $+ 0.55$. If now we combine the five variables having the highest coefficients of the 14th order (see table No.

3 col. No. 2) we find by computation that $R_1(2\ 4\ 6\ 8\ 9) = +0.52$. The difference between the two R 's is 0.03, probable error 0.09 (roughly), and therefore not only is not significant, but the chances are 5:1 in favor of its being due to chance. These five variables therefore contain everything that is significantly related to Marks, including the factor we are looking for. So if we eliminate successively No. 2 (Log. Mem. Aud.), No. 4 (Loss or Gain in No. 2), No. 6 (Sentences Built), No. 8 (Const. Increment), and No. 9 (Objects Seen), from r_{18} our problem will be solved.

Table No. 5 contains these values (see below). Inspection of it shows at once that r_{18} drops, as before, when No. 6 is eliminated, but rises again when No. 9 is eliminated. The factor we are looking for is common to No. 8 and No. 9, and is not in No. 2, No. 4, or No. 6.

TABLE No. 5

r_{18}	$= +0.21$
$r_{18 \cdot 2}$	$= +0.21$
$r_{18 \cdot 2\ 4}$	$= +0.21$
$r_{18 \cdot 2\ 4\ 6}$	$= +0.14$
$r_{18 \cdot 2\ 4\ 6\ 9}$	$= +0.19$

What is this factor? Well, I regret to say that we may save ourselves the trouble of interpretation, for the difference we have investigated is not based on a sufficient number of observations to be "significant." I have isolated this "factor" in order to exhibit what appears to me as the beauty of the analysis and in order to illustrate what can be done by means of indirect analysis, (analysis by exclusion) by means of the manipulation of R , a method, which, so far as I know, has not been suggested elsewhere.

Similar remarks are in order for tests No. 6 and No. 9. In both cases the correlation of the 14th order is significant. In both cases we have a fairly high degree of probability that each test possesses something, peculiar to it alone, significantly correlated with marks. Beyond this point our 92 observations will not permit us to go *statistically*. We may go further, if we like, by way of a priori reasoning which has perhaps sug-

gestive value. Moreover this value is enhanced by the fact that the elimination of fourteen fairly diverse tests limits us to suggestions which must be fairly concrete and—which is more to the point—verifiable by future investigations. Before the close of this paper I shall be guilty of a little speculation of this kind.

So far we have been occupied exclusively with analysis. To be sure, the paucity of our data, combined with the low value of our correlations, did not permit us to go very far. But I trust that the sort of thing which might be attained with more extensive data is clear. We may now turn to the subject of diagnosis and prognosis.

We have seen just now that five tests carry practically all the meaning, with reference to marks, contained in the fifteen tests. It follows that they carry also all the diagnostic value. We have shown in an earlier part of this paper that the value of R which may arise owing to fluctuations of sampling alone may easily become unpleasantly large. We found that for 16 variables (15 tests and a criterion) and 92 observations the “probable” value of R is $+ 0.40$ even when none of the tests have any actual relation to the criterion or to each other. Our result of $+ 0.55$, compared to this, has but little significance, and this is in itself a sufficient reason why it has little value for diagnosis. (I dare not say *no value*, on account of the argument we hear so often that an infinitesimal part of a loaf is better than no bread at all.) The case is more favorable if we combine our five “best” tests. Here the actual R is $+ 0.52$, and the corresponding “chance” value is $+ 0.21$. The difference, at any rate, is significant, and, if anyone cares to do it, he may compute the weights which these five tests have in their regression equation and use the equation for the “practical” diagnosis of the ability of individual students.²⁴ But considerations of a more familiar sort show us the trivial nature of the diagnostic value, even in this case, in still more drastic fashion. Let us assume that our $R = + 0.55$ not only is significant, but is *absolutely* correct for the entire “population” of freshmen in America. What then would be its diagnostic value? The

²⁴ A short method for computing weights will be found in the appendix.

probable error of R is now zero, but the "Standard Error of Estimate" remains. In discussing the logic of partial correlation we brought out the significance of this expression as being a "Residual." Now such a Residual, if it remains on our hands after we have estimated on the basis of all the information at hand,—be that information supplied by one or by fifty tests,—such a Residual is an Error of Estimate due to unknown causes. The standard deviation of all such errors is therefore a measure of the accuracy attainable on the basis of such information. In terms of the standard deviation of the variable we are estimating (say academic marks), it is $\sqrt{1 - r^2}$. As r approaches zero, $\sqrt{1 - r^2}$ approaches 1. Or, in other words, all of our estimates tend to coincide at the mean. Or, in still other words, in the absence of relation our average error of estimate will be least if we estimate all deviations as zero. Now when $r = 0.55$, $\sqrt{1 - r^2} = 0.84$. The gain in diagnostic value is thus very slight.²⁵

The case now stands as follows. The values of R , which we found, had no significance for a combination of fifteen tests, and only a moderate significance for a combination of five tests. But even if these values were absolutely correct, their diagnostic value would be very low. On the other hand our analysis of the Logical Memory test retains such significance as we showed it to have and, even though the lack of material did not permit of pushing the statistical analysis very far, it has focussed our attention on five of the tests in a way which could not have been anticipated from the raw material, and has, in some measure, cleared the way for further investigations.

Is this then the last word that can be said for "practical" diagnosis? By no means. To be sure, if diagnosis of individual ability be the end in view, the material we have investigated is worthless. Furthermore, I believe that this is true of all similar work done by others, in so far as their results can be duplicated

²⁵ The formula for the error of estimate is, of course, perfectly familiar. But, whatever may be true of others, I had never realized how very rapidly it increases in value as r becomes less than unity. My attention was called to the fact by Mr. B. Ruml.

by other investigators. But then, if diagnosis of *individual* ability for the purpose of educational guidance of the individual student were the only use to which tests and correlations could be put, I would not give the subject five seconds of my time. The motivation of this position would lead us too far afield, and we may waive discussion. But, the one who really needs guidance is the educator. And, even if the reader cannot agree with the first point, he will surely assent to the second. Now if it could be shown that a purely verbal type of memory has a correlation of zero with academic achievement at one institution or in one department, 25 at another, 50 at still another, etc., it would throw very little light on the "General Intelligence" of any individual student, but it would furnish a world of "guidance" to the educator. For, assuming that it is undesirable to encourage and develop the "verbal" type, he would be able to direct his attention and energy to the task of making other institutions or departments conform to the standard set by the first. It is to facts in the mass that statistics properly applies. It is there it can and should be applied.

IV. CONCLUSIONS BASED ON THE DATA

It remains to regale the reader with the *a priori* speculation with which he has been threatened. We may conveniently do this by casting it into the form of "conclusions." At the same time we will restate the more valid conclusions which have resulted from the discussion.

(1) The whole collection of tests has a low diagnostic value.

(2) This is due only in part to the low value of the individual correlations. It is due very largely to the enormous amount of duplication. The fifteen tests are, to a large extent, all measures the same thing or things. This fact enables us to concentrate our attention on five tests.

(Hereafter, in discussing the correlations of the various tests, reference will be to the coefficients of the highest order, unless specifically stated otherwise.)

(3) The Logical Memory Test: The probability is 1300:1 that test No. 2 is significant. The probability is 23:1 that audi-

tory presentation is "superior" to visual presentation. We need not repeat the discussion. We must not however interpret it to the effect that the "visual" test is of no value. It becomes superfluous when auditory presentation is used.

The probability is 140:1 that test No. 4 (loss or gain) is significant. This is a clue toward the differentiation of different types of memory. Comparison of various modes of scoring suggests itself. *E.g.*, we may compare, in No. 2, an evaluation based simply on the number of words correctly repeated with other methods. The significance of No. 4 may also be due to the factor of "interest." To a lesser degree of probability, similar statements can be made about test No. 5.

(4) The probability is about 25:1 that test No. 8 (Const. Increment) is significant. The instructions for this test emphasize speed more than those for any of the others where speed is a factor at all. Moreover the activity is far more mechanical and automatic, approximating simple reaction time after practice. Indeed for the faster subjects the limit of speed seems to be physiological rather than mental. Even some subjects who are only fairly fast give this impression, *i.e.*, it looks as if they cannot talk as fast as they can add. This, physiological reaction time, may be the significant factor. It might be investigated by using 1 as the increment, adding only to single digits, and reducing the number to fifty. On the other hand the test, as given, is monotonous and long. It is hard work, and fast subjects are a little out of breath when they finish. As has been said, the raw data were not available, but I have noted elsewhere that some subjects increase their speed as they go along and spurt at the end, while others become slower and slower. By contrasting speed say in four quarters we might get a measure of some of the so-called character qualities, such as effort, perseverance, etc. Some such thing may be the significant factor in the Constant Increment Test. Of course, all this is highly speculative.

(5) The probability is about 30:1 that test No. 9 (Objects Seen) has *negative* significance. Of the ten objects presented, the fountain-pen, penfiller and inkwell, the envelope and the two cent stamp, the pencil and the ruler, are fairly well associated.

The association for the 25 cent piece, the maroon ribbon, and the key are more remote. The mean for all subjects was a little over seven, so that those who recalled more than the number of objects which are closely associated tended to have low marks. Now it is well known that the capacity to note and to retain a mass of irrelevant details is very great in the hypnotic trance, and many authorities incline to the belief that the trance is simply a state of diffuse attention. It might very well be that a selective, discriminating type of mind will obtain low scores in this test. The suggestion is of course capable of experimental investigation by varying the nature of the objects.

(6) There is a probability of about 30:1 that test No. 6 (Sentences Built) is significant. I have no comment to offer.

(7) Accuracy appears to be of no significance except possibly as it enters into the logical memory tests. In the case No. 7 (Hard Directions) there is even a probability of 5:1 that it has negative significance. Now it is fairly well established that speed and accuracy are positively correlated at many activities and this is also born out by our data. Thus $r_{8\ 15} = +0.40$ (Const. Inc.), $r_{7\ 16} = +0.25$ (H. Directions) and $r_{11\ 14} = +0.18$ (Opposites). On the other hand the relation is probably inverse, within limits, for any given individual. Of course, our tests cannot show the deviation of an individual from his own point of maximum efficiency. So it would seem that in so far as speed is significant, it has associated with it all of the significance of accuracy. We have already seen that in so far as speed is significant, it seems to be most adequately represented by the Constant Increment test.

If it is hard to say why a test does correlate, it is even more difficult to indicate why it does not. We may say in a general way that the ten tests whose coefficients of the highest order give no indication of significance, are very likely influenced by general factors, such as ability to adapt to novel situations such as the entire test situation is for the average subject, by interest, effort, etc. We can only say that in so far as they have value they are more adequately represented by the five tests with the highest coefficients of the 14th order.

This is all that I find it useful to say of the specific results obtained in the present study. We will now turn to an exposition of the mechanical technique involved in analyzing a complex situation.

V. APPENDIX: A CONTRIBUTION OF THE MECHANICAL TECHNIQUE OF PARTIAL CORRELATION

In what follows it is assumed that the reader already has a working knowledge of the theory of correlation. To proceed on any other assumption would be impossible in a work of the scope of the present one. On the other hand it has been my aim to arrange the steps in such a way that any-one who understands the terminology can follow the steps mechanically, or at least by symmetry, without having to inquire about the why's and wherefore's.

The general problem before us is this: Given a dependent variable and a number of independent ones, how can we obtain a maximum of information with a minimum of arithmetic. In so far as the schema devised by me is applicable to such a problem it can be divided into three sub-problems, which however form successive stages of what really is a single operation. These three stages are: (1) The finding of the coefficient of Multiple correlation. (2) The finding of the coefficients of partial correlation, with reference to the dependent variable, of the highest order. (3) The finding of the coefficients of regression (weights) for estimating the dependent variable.

For the sake of simplicity we will assume that all standard deviations of zero order are 1. Also, in order to avoid the cumbersome use of "n," let us suppose that we are dealing with five independent variables, and one dependent one. The method can be extended to any number whatever by symmetry.

PROBLEM I

Find $R_1(23456)$. This may be done directly from the equation²⁶

$$1-R^2_{1(23456)} = (1-r^2_{12})(1-r^2_{13.2})(1-r^2_{14.23})(1-r^2_{15.234})(1-r^2_{16.2345}) \quad \text{No. 1}$$

Beginning with the raw data, the work proceeds as follows:

(1) Find all coefficients of zero order.

²⁶ Yule. Intro., p. 248.

- (2) Write equation No. 1 as above, and rewrite it, reversing the order of subscripts as follows:

$$1-R^2 = (1-r^2_{16})(1-r^2_{15.6})(1-r^2_{14.56})(1-r^2_{13.456})(1-r^2_{12.3456}) \quad \text{No. 2}$$

- (3) Compute the coefficients of equation No. 1 and No. 2. For this purpose there will be needed all coefficients having the secondary subscripts in these two equations, forty all told. That is, for equation No. 1, we will need all coefficients of type r_{-2} , r_{-23} , r_{-234} , r_{-2345} . It is convenient to prepare a list of these coefficients on which their values can be entered as fast as computed. Such a list can be written by symmetry with a minimum of thought, and about as fast as one can write. Giving subscripts only, it is:

TABLE I

13.2				15.6			
14.2	14.23			14.6	14.56		
15.2	15.23	15.234		13.6	13.56	13.456	
16.2	16.23	16.234	16.2345	12.6	12.56	12.456	12.3456
34.2				25.6			
35.2				24.6	24.56		
36.2				23.6	23.56	23.456	
45.2	45.23			35.6			
46.2	46.23			34.6	34.56		
56.2	56.23	56.234		45.6			

- (4) Compute $1-R^2$ according to equations No. 1 and No. 2. The two values should check. They afford an independent check on all of the previous arithmetic with exception of the computation of the coefficients of zero order.
- (5) Compute R , or, preferably, look it up.²⁷

PROBLEM II

To find the coefficients of the fourth order with reference to the dependent variable. There are 5 such coefficients. They are:

$r_{12.3456}$

$r_{13.2456}$

$r_{14.2356}$

$r_{15.2346}$

$r_{16.2345}$

²⁷ Use "Tables for statisticians and Biometricians." K. Pearson, Table 8, pp. 20-21.

$r_{12 \cdot 3456}$ and $r_{16 \cdot 2345}$ have been found in Problem No. 1. The other three can be found indirectly from the following consideration. Let $R_1(-_a)$ be the Multiple coefficient showing the relation of X_1 to a combination of all the independent variables except X_a , *e.g.*, $R_1(-_2)$ is $R_1(3456)$.

No. 3

$$\text{then } \frac{1 - R^2_{1(abc \cdot n)}}{1 - R^2_{1(-_a)}} = 1 - r^2_{1a \cdot bc \cdot \cdot n}$$

$$\text{so that } \frac{1 - R^2_{1(23456)}}{1 - R^2_{1(2456)}} = 1 - r^2_{13 \cdot 2456}$$

$$\frac{1 - R^2_{1(23456)}}{1 - R^2_{1(2356)}} = 1 - r^2_{14 \cdot 2356}$$

$$\frac{1 - R^2_{1(23456)}}{1 - R^2_{1(2346)}} = 1 - r^2_{15 \cdot 2346}$$

We may write,

$$\begin{aligned} 1 - R^2_{1(2456)} &= (1 - r^2_{16})(1 - r^2_{15 \cdot 6})(1 - r^2_{14 \cdot 56})(1 - r^2_{12 \cdot 456}) \\ 1 - R^2_{1(2356)} &= (1 - r^2_{16})(1 - r^2_{15 \cdot 6})(1 - r^2_{13 \cdot 56})(1 - r^2_{12 \cdot 356}) \\ 1 - R^2_{1(2346)} &= (1 - r^2_{12})(1 - r^2_{13 \cdot 2})(1 - r^2_{14 \cdot 23})(1 - r^2_{16 \cdot 234}) \end{aligned}$$

Now all of these coefficients are in table I with exception of $r_{12 \cdot 356}$, which may be computed from $r_{12 \cdot 56}$, $r_{13 \cdot 56}$, and $r_{23 \cdot 56}$ which also are in table I. We are now able to compute all of the coefficients of the fourth order from equations of type No. 3.

It will be noted that coefficients found indirectly from equation No. 3 are indeterminate as to sign. However, unless the numerical magnitude of such coefficients is negligible, it will usually be possible to determine the sign by inspection.

PROBLEM III

Find the coefficients of regression. The familiar expression

$$\text{for this coefficient is, } b_{12 \cdot 34 \cdot \cdot n} = r_{12 \cdot 34 \cdot \cdot n} \frac{\sigma_{1 \cdot (234 \cdot \cdot n)}}{\sigma_{2 \cdot 134 \cdot \cdot n}}$$

We need therefore, in addition to the correlation coefficients of the fourth order, all of the six standard deviations of the fifth order. These should be written as follows:

$$\sigma_{1 \cdot 23456} = \left[(1-r^2)_{16} (1-r^2)_{15.6} (1-r^2)_{14.56} (1-r^2)_{13.456} (1-r^2)_{12.3456} \right]^{\frac{1}{2}}$$

$$\sigma_{2 \cdot 13456} = \left[(1-r^2)_{26} (1-r^2)_{25.6} (1-r^2)_{24.56} (1-r^2)_{23.456} (1-r^2)_{12.3456} \right]^{\frac{1}{2}}$$

$$\sigma_{3 \cdot 12456} = \left[(1-r^2)_{36} (1-r^2)_{35.6} (1-r^2)_{34.56} (1-r^2)_{23.456} (1-r^2)_{13.2456} \right]^{\frac{1}{2}}$$

$$\sigma_{4 \cdot 12356} = \left[(1-r^2)_{46} (1-r^2)_{45.6} (1-r^2)_{34.56} (1-r^2)_{24.356} (1-r^2)_{14.2356} \right]^{\frac{1}{2}}$$

$$\sigma_{5 \cdot 12346} = \left[(1-r^2)_{25} (1-r^2)_{35.2} (1-r^2)_{45.23} (1-r^2)_{56.234} (1-r^2)_{15.2346} \right]^{\frac{1}{2}}$$

$$\sigma_{6 \cdot 12345} = \left[(1-r^2)_{26} (1-r^2)_{36.2} (1-r^2)_{46.23} (1-r^2)_{56.234} (1-r^2)_{16.2345} \right]^{\frac{1}{2}}$$

The only new coefficient in all of the above expressions is $r_{24 \cdot 356}$. It can be found from $r_{23 \cdot 56}$, $r_{24 \cdot 56}$, and $r_{34 \cdot 56}$, table I. The above six expressions for the standard deviations of the fifth order are written in such a way that the dependent variable will enter into the last term as a primary subscript, and from this point the variables are eliminated as nearly as possible in the same order in which they have been eliminated in equations No. 1 and No. 2, choosing by inspection the equation which is seen to be best for the purpose.

From this point on the computation of the regression coefficient proceeds in the usual way.

The method outlined will, of course, lend itself to the solution of a variety of problems. The only point at all novel is expressed by equation No. 3, problem No. 2. In spite of its very great simplicity, I have never seen it in print, probably because it has no theoretical interest. It does however afford a very useful shortcut to the arithmetic, and the rest of the schema is simply a systematic exploitation of this fact.

In comparing this schema with the one given in Yule's "Introduction," it should be borne in mind that Yule's schema provides for the finding of all possible relations. So far as I know, there is no short cut to this problem. I have simply taken advantage of the fact that in "test" work interest centers, or should center, on one or two dependent variables, not themselves tests.

The schema will show to advantage, however, if contrasted against's Kelley's.²⁸ Kelley faces practically the same problem as I, except that he scarcely mentions analysis and puts the emphasis exclusively on diagnosis and prognosis. He states²⁹ that the number of coefficients of partial correlation to be computed is 2 in the case of three variables, 15 in the case of 4 variables, 36 in the case of 5 variables, 78 in the case of 6 variables, etc. Our schema requires 2 coefficients in the case of 3 variables, 8 in the case of 4 variables, 22 in the case of 5 variables, 45 in the case of 6 variables, etc. Moreover, in our method, the symmetry of table I, which accomplishes practically all of the work, is very easily seen and reproduced, and it requires only a very moderate amount of practice and ingenuity to write the expressions in problems No. 1 and No. 2 to the best advantage. Over against this, Kelley does not state any guiding principle for selecting the coefficients to be computed to the best advantage and I am quite unable to see the symmetry without such aid. For example I am quite unable to say how many coefficients Kelley *would* need for say 8 variables. But then I am quite prepared to find that that is due to my own stupidity.

Finally I wish to call attention again to the very useful character of Kelley's tables. Using these tables in conjunction with a "millionaire" calculating machine, 75 coefficients per hour, correct to two places, or 40 coefficients, correct to three places, can easily be computed.

²⁸ *Op. cit.*, p. 23, this paper.

²⁹ *Op. cit.*, p. 14.

PSYCHOLOGICAL MONOGRAPHS

VOL. XVII

72. *An Experimental Study of Decision Types and their Mental Correlates.* JAMES W. BRIDGES. Pp. 72. 75 cents. 73. *The Genetic Aspect of Consonance and Dissonance.* HENRY T. MOORE. Pp. 68. 75 cents. 74. *The Influence of Distractions of the Formation of Judgments in Lifted Weight Experiments.* DAVID MITCHELL. Pp. 58. 50 cents. 75. *Yale Psychological Studies, New Series, Vol. II, No. 1.* Edited by ROSWELL P. ANGLIER. Pp. 155. \$1.75. 76. *The Measurement of Attention.* HERBERT WOODROW. Pp. 158. \$1.50.

VOL. XVIII

77. *Mental and Physical Measurements of Working Children.* HELEN T. WOOLLEY and CHARLOTTE R. FISHER. Pp. 247. \$2.50. 78. *Recognition and Discrimination.* GUSTAVE A. FEINGOLD. Pp. 128. \$1.25. 79. *Alternation and Interference of Feelings.* CHESTER ELIJAH KELLOGG. Pp. 94. \$1.00. 80. *A Study in Association Reaction and Reaction Time.* HARRY W. CRANE. Pp. 75. 75 cents.

VOL. XIX

81. *I. Symptomatological Differences Associated with Similar Cerebral Lesions in the Insane. II. Variations in Distribution of the Motor Centers.* SHEPHERD IVORY FRANZ. Pp. 162. \$1.50. 82. *The Psycho-physiological Effect of the Elements of Speech in Relation to Poetry.* ROBERT C. GIVLER. Pp. 132. \$1.25. 83. *Standardization of Tests for Defective Children.* CLARA SCHMITT. Pp. 181. \$1.75. 84. *A Study of Retroactive Inhibition.* J. EDGAR DE CAMP. Pp. 69. 75 cents.

VOL. XX

85. *A Horizontal-Vertical Illusion of Brightness in Foveal Vision Apparent in Astronomical Observations of the Relative Luminosity of Twin Stars.* JOSEPH W. HAYES. Pp. 126. \$1.25. 86. *Recognition: A Logical and Experimental Study.* ROBERT B. OWEN. Pp. 154. \$1.50. 87. *Formal Discipline from the Standpoint of Experimental Psychology.* JOHN EDGAR COOVER. Pp. 307. \$3.00. 88. *Learning Tests with Deaf Children.* RUDOLPH PINTNER and DONALD G. PATERSON. Pp. 58. 75 cents.

VOL. XXI

89. *Mental Measurements of the Blind.* THOMAS H. HAINES. Pp. 86. \$1.00. 90. *The Process of Generalizing Abstraction; and its Product, the General Concept.* SARA CAROLYN FISHER. Pp. 212. \$2.00. 91. *Acquisition of Skill.* WM. H. BATSON. Pp. 92. \$1.00. 92. *Studies in Social and General Psychology from the University of Illinois.* Edited by MADISON BENTLEY. Pp. 115. \$1.00.

VOL. XXII

93. *Voluntary Isolation of Control in a Natural Muscle Group.* J. C. BARNES. Pp. 50. 50 cents. 94. *Psycho-Motor Norms for Practical Diagnosis.* J. E. W. WALLIN. Pp. 101. \$1.00. 95. *Apparatus and Experiments on Sound Intensity.* A. P. WEISS. Pp. 59. 75 cents. 96. *Wellesley College Studies in Psychology Number 2.* Edited by ELEANOR A. MCC. GAMBLE. Pp. 191. \$1.75. 97. *Children's Association Frequency Tests.* HERBERT WOODROW and FRANCES LOWELL. Pp. 110. \$1.25.

VOL. XXIII

98. *Scientific Study of the College Student.* HARRY DEXTER KITSON. Pp. 81. 75 cents. 99. *Whole vs. Part Methods in Motor Learning. A Comparative Study.* LOUIS AUGUSTUS PECHSTEIN. Pp. 88. 75 cents. 100. *Yale Psychological Studies, New Series, Vol. II, No. 2.* Edited by ROSWELL P. ANGLIER. Pp. 150-331. \$1.75. 101. *Studies from the Psychological Laboratory of the University of Chicago. The Vertical-Horizontal Illusion.* SARAH MARGARET RITTER. Pp. 114. \$1.25.

VOL. XXIV

102. *Two Studies in Mental Tests. I. Variable Factors in the Binet Tests. II. The Diagnostic Value of Some Mental Tests.* CARL C. BRIGHAM. Pp. 254. \$2.50. 103. *Radiometric Apparatus for Use in Psychological and Physiological Optics.* C. E. FERREE and GERTRUDE RAND. Pp. xvi+ 65. 75 cents.

Directory of American Psychological Periodicals

American Journal of Psychology—Worcester, Mass.: Florence Chandler.
Subscription \$5. 600 pages annually. Edited by G. Stanley Hall.
Quarterly. General and experimental psychology. Founded 1887.

Pedagogical Seminary—Worcester, Mass.: Florence Chandler.
Subscription \$5. 575 pages annually. Edited by G. Stanley Hall.
Quarterly. Pedagogy and educational psychology. Founded 1891.

Psychological Review—Princeton, N. J.: Psychological Review Company.
Subscription (with Psychological Bulletin) \$6. 480 pages annually.
Bi-monthly. General. Founded 1894. Edited by Howard C. Warren.

Psychological Bulletin—Princeton, N. J.: Psychological Review Company.
Subscription \$3.25. 480 pages annually. Psychological literature.
Monthly. Founded 1904. Edited by Shepherd I. Franz.

Psychological Monographs—Princeton, N. J.: Psychological Review Company.
Subscription \$4. per vol. 500 pp. Founded 1895. Ed. by James R. Angell.
Published without fixed dates, each issue one or more researches.

Psychological Index—Princeton, N. J.: Psychological Review Company.
Subscription \$1.25. 200 pp. Founded 1895. Edited by Madison Bentley.
An annual bibliography of psychological literature.

Journal of Philosophy, Psychology and Scientific Methods—New York:
Sub-Station 84. Bi-weekly. 728 pages per volume. Founded 1904.
Subscription \$3. Edited by F. J. E. Woodbridge and Wendell T. Bush.

Archives of Psychology—Sub-station 84, N. Y.: Archives of Psychology.
Subscription \$5. 600 pp. ann. Founded 1906. Ed. by R. S. Woodworth.
Published without fixed dates, each number a single experimental study.

Journal of Abnormal Psychology—Boston; Richard G. Badger.
Subscription \$4. 480 pages annually. Edited by Morton Prince.
Bi-monthly. Founded 1906. Entire field of abnormal psychology.

Psychological Clinic—Philadelphia: Psychological Clinic Press.
Subscription \$1.50. 280 pages annually. Edited by Lightner Witmer.
Monthly (9 numbers). Orthogenics, psychology, hygiene. Founded 1907.

Training School Bulletin—Vineland, N. J.: The Training School.
Subscription \$1. 160 pp. ann. Ed. by E. R. Johnstone. Founded 1904.
Monthly (10 numbers). Psychology and training of defectives.

Journal of Religious Psychology—Worcester, Mass.: Louis N. Wilson.
Subscription \$3. 480 pages per vol. Founded 1904. Ed. by G. Stanley Hall.
Published without fixed dates, 4 numbers constitute a volume.

Journal of Race Development—Worcester, Mass.: Louis N. Wilson.
Subscription \$2. 460 pages annually. Founded 1910.
Quarterly. Edited by George H. Blakeslee and G. Stanley Hall.

Journal of Educational Psychology—Baltimore: Warwick & York.
Subscription \$2.50. 600 pages annually. Founded 1910.
Monthly (10 numbers). Managing Editor, J. Carleton Bell.
(Educational Psychology Monographs. Edited by Guy M. Whipple.
Published separately at varying prices. Same publishers.)

Journal of Animal Behavior—Cambridge, Mass.: Emerson Hall.
Subscription \$5. 450 pp. annually. Founded 1911.
Bi-monthly. Robert M. Yerkes, Managing Editor.

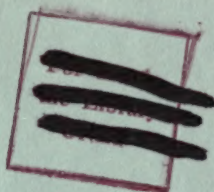
The Behavior Monographs—Cambridge, Mass.: Emerson Hall.
Subscription \$5. 450 pages per volume. Edited by John B. Watson.
Published without fixed dates, each number a single research.

Psychoanalytic Review—New York: 64 West 56th Street.
Subscription \$5. 500 pages annually. Psychoanalysis.
Quarterly. Founded 1913. Ed. by W. A. White and S. E. Jelliffe.

Journal of Experimental Psychology—Princeton, N. J.: Psychological Review Company. 480 pages annually. Experimental. Founded 1916.
Subscription \$3.25. Bi-monthly. Edited by John B. Watson.

BF
1
P69
v.24

Psychological monographs: gen-
eral and applied



CIRCULATE AS MONOGRAPH



**PLEASE DO NOT REMOVE
SLIPS FROM THIS POCKET**

**UNIVERSITY OF TORONTO
LIBRARY**

